

Station at a crossroads

Frank international discussions need to start immediately if anything is to be salvaged from the space station, whose completion currently relies on the ailing space shuttle.

Until about a week ago, most observers of the space shuttle assumed that the fleet could be kept alive until its planned retirement in 2010.

But the latest mission of the shuttle Discovery, with its daily litany of stuck fuel gauges, falling foam and chipped tiles (see page 608), raises the prospect that this cannot happen. The ageing shuttle's problems may soon become so difficult to analyse and so expensive to fix that even its staunchest defenders will see that the time has come to stop throwing good money after bad.

What then? The International Space Station is at least 15 shuttle flights away from completion, and that's just counting its largest elements, the European and Japanese laboratory modules and the trusses for solar-power arrays. Several more flights are needed to haul up the experimental racks that would equip the laboratories.

The Russian Soyuz crew vehicle and Progress supply ship are each far too small to carry these large components into orbit. Europe's Automated Transfer Vehicle (ATV), a new cargo carrier scheduled to debut on an Ariane rocket next year, can deliver tons of supplies but not large sections of the station. Japan's proposed cargo carrier, called the HTV, can handle the experiment racks, but won't enter service until 2010. So abandoning the shuttle now would leave the station in its current, half-finished state.

One alternative to that would be for NASA to start work as quickly as possible on a shuttle-derived vehicle (SDV) that would replace the component of the shuttle that carries astronauts with a giant cargo pod. Such an approach is needed anyway for the proposed return to the Moon. In principle, the SDV could deliver the rest of the large pieces of the station, which astronauts, ferried to space on Russian vehicles, could assemble in orbit. In the four or five years it would be expected to take to design and build the SDV, Russian vehicles and Europe's ATV could keep the station aloft, lightly staffed and stocked.

Such a plan would require Europe and Japan to accept yet another major delay to the date on which their labs will enter operation.

They have already stood by helplessly for years, watching NASA make essentially unilateral decisions to scale back the design according to the vagaries of US budget politics. Why should they continue to put up with this?

Well, for one thing, they may not have much choice. But it also runs counter to the interests of the Japanese and European space agencies to watch NASA — which leads most international space projects and will continue to do so for the foreseeable future — damage its reputation in further, forlorn efforts to keep the ailing shuttle in space.

The international partners could also make use of the delay to negotiate better terms for their participation in the space-station project. Michael Griffin, the latest NASA administrator, has made no secret of his low regard for the station since his appointment in April. Except for medical research on astronauts and technology tests related to the Moon-Mars programme, NASA's use of it is likely to be scaled back, so it ought to cede more laboratory time, more astronaut participation and more mission-control involvement to Europe and Japan.

Such an arrangement would assume that NASA's long-suffering international partners would relish an enhanced role in the project. Publicly, their commitment to the space station is as robust as ever. But if, in truth, they'd rather leave the project in its current state, abandon their laboratory modules, and start spending their taxpayers' money on something more useful, now is surely the time to say so. NASA could then offer something else — probably a prominent role in other cooperative projects — to compensate for reneging on its obligation to complete the US end of the deal. Either way, it's time for some plain speaking and creative thinking, not for stubbornly sticking to an obsolete plan. ■

"It runs counter to the interests of Japan and Europe to watch NASA damage its reputation in further forlorn efforts to keep the shuttle in space."

Count themselves lucky

Mathematicians might think they have an image problem, but the public holds them in great esteem.

Like people in many disciplines, mathematicians are prone to bouts of concern that they have an image problem. Only last month, some of them convened on the Greek island of Mykonos with a group of writers to consider how a better use of narrative could help them with their work — and with their public relations (see page 622).

It is probably fair to say that many mathematicians feel themselves perceived as unable to conduct the simplest practical task, unfashionably attired, nerdy and isolated from the real world.

A collection of the jokes that mathematicians tell each other (*Not. Am. Math. Soc.* 52, 24; 2005) reveals an element of self-mockery of their obsessive and pedantic natures. Who else would laugh at the suggestion that what you get by crossing an elephant with a banana is $|\text{elephant}| * |\text{banana}| * \sin \theta$?

Additionally, as they are never shy to tell the rest of us, mathematicians receive only a tiny slice of the overall research funding. And although it clearly costs much less to prove a theorem than it does to clone a cow, their small grants are inevitably interpreted



by mathematicians as a sure sign that their work is undervalued.

These negative associations have been reinforced by a number of popular stories about great mathematicians. The American inventor of cybernetics theory, Norbert Wiener, for example, is frequently depicted as the archetypal absent-minded professor. It is said he once lost his way walking home from the Massachusetts Institute of Technology. He came across a small girl in the street and asked if she could give him directions. "Yes, daddy," she replied, "I'll take you home."

Kurt Gödel, whose incompleteness theorem sent shock waves through mathematics in the 1930s, was a noted misanthrope, who shunned human contact at the Institute for Advanced Study in Princeton, preferring colleagues to communicate via pieces of paper stuffed through the crack beneath the door of his office.

Despite — or, perhaps, because of — such behaviour, the history of mathematics is probably more colourful than that of any other scientific discipline. And there seems to be an insatiable public appetite for tales of the almost supernatural intellectual powers of its more famous figures.

Srinivasa Ramanujan, for example, an Indian mathematician of towering ability in number theory who died at the age of only 32, first came to the attention of the eminent British mathematician G. H. Hardy by sending his notebooks to him while he worked as a clerk in Madras. Hardy correctly concluded that even if he couldn't follow all of the proofs, only a genius could have thought of the theorems they were seeking to address.

Hardy invited him to Cambridge, but Ramanujan caught a cold that developed into a terminal case of tuberculosis. When Hardy visited his ailing protégé one day by taxi, he commented that the cab's number, 1729, was "rather dull". On the contrary, Ramanujan insisted, it is the smallest number expressible as the sum of two different pairs of cubes.

Earlier eras have produced equally poignant anecdotes. One thinks, for example, of Évariste Galois, the unruly French mathematician who made great strides in group theory. He frantically scribbled down his work for posterity on the eve of his fatal duel in 1832 at the age of just 20. Such stories have propelled books such as Simon Singh's on Fermat's last theorem to bestseller status.

These tales are popular not just for their panache, but because they celebrate mathematicians as pure intellectuals who, unlike physicists, biologists or chemists, are untainted by applications of their work. For even though mathematics is eminently useful, its application barely features in its public reputation. Disciplines that are traditionally inclined to disdain pure theory — biology springs to mind — should take note of the success with which mathematics, this most theoretical of disciplines, has haplessly bungled its way into people's hearts. ■

"There seems to be an insatiable public appetite for tales of the almost supernatural intellectual powers of mathematics' more famous figures."

A dog's life

The first cloned dog was born at some cost, and there needn't be many more.

An Afghan hound born in South Korea in June adds dogs to the small list of animal species that have been successfully cloned (see page 641). The birth marks another first for the Korean-based group that cloned the first human embryos last year.

The development has some scientific significance, on account of the emerging importance of the dog as a model for the study of certain aspects of human genetics, development, behaviour and disease.

A dog genome project is being undertaken by a US team, and the cloning of dogs could provide an additional tool for researchers. The number of cloned dogs that will be needed for such research is probably small, however. Scientists such as Elaine Ostrander of the US National Human Genome Research Institute, head of the dog-genome project, do most of their work with pets living at home, not with kennels of animals bred for research. So the ability to clone dogs is unlikely to have more than a marginal impact on how such research is done.

Cohorts of cloned dogs could potentially be used to study the respective influence of genes and environment on particular traits, however. And if it were possible to derive embryonic stem-cell lines from cloned dog embryos — something that's so far only been done in mice and humans — then canine diseases could be studied more easily in Petri dishes, perhaps providing insights into disease

mechanisms and even identifying new therapies. Deriving embryonic stem cells would also pave the way to therapeutic cloning in dogs — perhaps providing a useful animal model for research into human health.

The initial dog-cloning experiment has proven the process to be remarkably inefficient, however, with only two live births — and one survivor — from a total of 1,095 embryos implanted in 123 surrogate mothers. This offers scant prospects for commercial pet cloning, the application of the work that the media is likely to make a fuss about. It is unlikely that even the most obsessive pet owner would contemplate preparing more than 100 failed pregnancies for just one successful birth — especially when there is no guarantee that the cloned dog will behave like the one they hope to duplicate. In such circumstances, the cloning of dogs for pet owners remains ethically indefensible.

The Korean researchers named their new dog Snuppy, for Seoul National University puppy (one can almost imagine the name being chosen — presumably on a conference call with the university press office). Let us wish him a long and happy life and hope that now that the concept behind the birth is proven, dogs are cloned only when strictly required for research purposes, and that effort is concentrated on work that carries the most likely rewards for canine and human health. ■

"It is unlikely that even the most obsessive pet owner would contemplate preparing more than 100 failed pregnancies for just one successful birth."

RESEARCH HIGHLIGHTS

A fish cooperative

Biol. Lett. doi:10.1098/rsbl.2005.0344 (2005)

A study of small 'cleaner' fish, which groom larger reef fish in return for protection, has shown the sacrifice they make to maintain this precarious cooperation.

The cleaner wrasse *Labroides dimidiatus* helps its host by gobbling parasites, but can also do damage by tucking into its host's tastier mucus. Observations have suggested that reef fish punish their cleaners for eating mucus either by chasing them or by finding a new cleaner.

Redouan Bshary from the University of Neuchâtel, Switzerland, and Alexandra Grutter from the University of Queensland, Australia, show that *L. dimidiatus* learns to avoid punishment by changing its feeding habits.



G. DOUWMA / NATUREPL.COM

CHEMICAL PHYSICS

Born into nobility

Europhys. Lett. **71**, 276–282 (2005)

Palladium has been endowed with noble status by Erwin Hümer of the Technical University of Clausthal and Krzysztof Osuch of the University of South Africa.

The noble behaviour, or relative lack of reactivity, of metals such as copper, silver and gold results from the complete filling of an energy level in the metal called the *d* band. Lone atoms of palladium have filled *d* shells, but in the bulk metal the band partly empties.

Hümer and Osuch deposited a layer of palladium on niobium, which pushes the atoms further apart, reasoning that this might restore electrons to the *d* band. The resulting monolayer was as unreactive as silver — partly because the *d* band became more nearly filled, but also because the band's energy was lowered, which made the electrons less available for reactions.

DEVELOPMENTAL BIOLOGY

Scent detectors

Science **309**, 787–790 (2005)

Caenorhabditis elegans, everyone's favourite model worm, has been shown to have a long-lasting memory for a smell associated with food, providing it is exposed to the cue during its first larval stage. Adult worms primed in their youth to respond to benzaldehyde, which smells a bit like marzipan, reacted to the odour by laying more eggs and by moving quickly towards its source.

Jean-Jacques Remy of the Developmental Biology Institute of Marseille, France, and Oliver Hobert of the Columbia University Medical Center in New York, also identified the protein required for this olfactory imprinting. Called SRA-11, it belongs to a class of olfactory receptors, but shows up on connecting interneurons rather than on sensory neurons. Its precise role is a mystery.

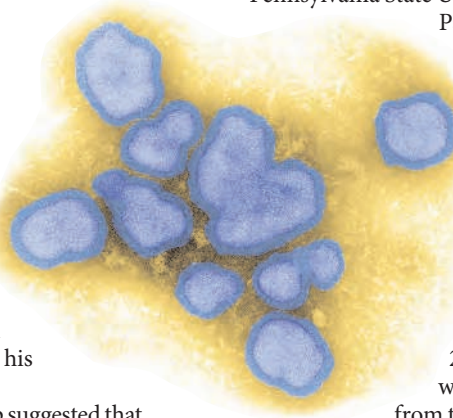
FERTILITY

Egg boxes

Cell **122**, 303–315 (2005)

Cells from the blood and bone marrow can restock female mammals' ovaries with eggs. This is the controversial claim of Jonathan Tilly of the Massachusetts General Hospital in Boston and his colleagues.

Last year Tilly's group suggested that mouse ovaries contain stem cells that produce new eggs in adulthood, challenging the dogma that female mammals are born with a fixed number of eggs. Now they have identified cells in bone marrow and blood that make proteins characteristic of germ cells. They have also shown that bone-marrow transplants or blood transfusions result in donor-derived eggs appearing in the ovaries of chemically sterilized female mice. However, the team has not demonstrated fertilization of these eggs or embryo development.



VIRAL GENETICS

Catching the flu

PLoS Biol. **3**, 300 (2005)

Influenza viruses (pictured) are known to swap genes. This process, called reassortment, may produce more virulent strains. To quantify the rate at which reassortment happens, Edward Holmes of Pennsylvania State University in College

Park and his colleagues studied the genomes of 156 H3N2

influenza A strains collected in New York state between 1999 and 2004. They found more gene swapping than expected, and showed that a flu epidemic during winter

2003–04 was caused when a dominant strain

from the previous year picked up a gene for a key surface protein from a less common strain.

BIOTECHNOLOGY

Safe delivery

Nature Biotechnol. doi: 10.1038/nbt1122 (2005)

Scientists hope to exploit the recently discovered class of molecules called small interfering RNAs, which target and shut down specific genes, as novel therapies. But a stumbling block is that introduced

SCIENCE SOURCE / SPL

RNAs are rapidly degraded in the body.

Now, researchers led by David Morrissey at Sirna Therapeutics in Boulder, Colorado, and at Protiva Therapeutics in Burnaby, British Columbia, have found a way to protect the RNA from destruction using a lipid bilayer.

They enclosed siRNAs directed against the hepatitis B virus within bilayer particles, and injected the particles into mice infected with hepatitis B. The siRNA inhibited viral replication for up to seven days at doses low enough to avoid toxicity.

MATERIALS

Supermarket sweep

Nature Mater. doi:10.1038/nmat1434 (2005)

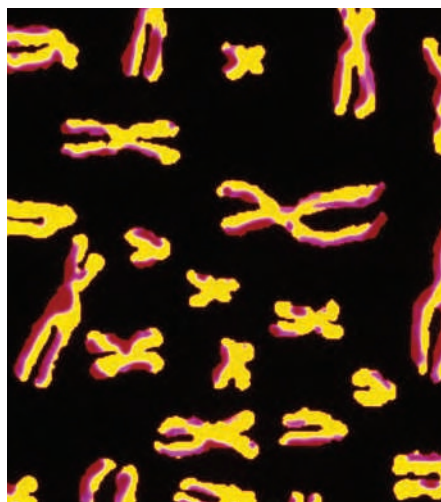
There is a potentially vast market for cheap radio frequency identification tags — which store information like barcodes, but can be read by radio scanners. This is driving efforts to make tag components from organic materials that could be printed directly on to packaging. A team from the University of Leuven, Belgium, has conquered one hurdle, manufacturing a diode from the organic compound pentacene that operates quickly enough to process current in tags read by high-frequency radio signals.

GENETICS

Generation X

Science **309**, 768–771 (2005)

In female mammals, each cell shuts off one of its two X chromosomes. But how cells know that they have multiple copies of the X is little understood. Now, Jeannie Lee of Harvard Medical School in Boston, Massachusetts, reports that the *Tsix* and *Xite* genes seem to house the 'counting' mechanism.



Female mouse cells with *Tsix* deletions randomly inactivate one or both of their X chromosomes or neither of them, whereas female cells with extra copies of *Tsix* and *Xite* fail to initiate inactivation.

DRUG DISCOVERY

Poor resistance

Proc. Natl Acad. Sci. USA

doi:10.1073/pnas.0504952102 (2005)

Compounds that inhibit the activity of kinase enzymes have been used to treat some kinds of cancer with dramatic success. But a drawback of their clinical introduction has been the rapid emergence of mutant drug-resistant kinases in treated patients.

A research team led by Patrick Zarrinkar and David Lockhart from Ambit in San Diego demonstrate a strategy that could help tackle this trend, which takes advantage of the tendency of kinase inhibitors to hit multiple targets. The team screened various kinase inhibitors for their effects on three drug-resistant kinases. Although not designed to act on these targets, some were effective against the mutant kinases. Therapeutic deployment could be quick because these compounds are already in clinical development.

OPTICS

A clearer view

Appl. Phys. Lett. **87**, 034102 (2005)

In principle, flat lenses that focus light to a point of infinite precision can be made using materials with a negative refractive index. Such materials, which can be made from arrays of loops of wire, bend light in the opposite direction to classical materials. But these 'metamaterials' also absorb much of the light's energy, thereby clouding the view through the lens. To counter this problem, Steven Anlage and his colleagues from the University of Maryland, College Park, built a metamaterial from superconducting niobium metals. Using this material led to significantly reduced absorption in the lens, and to better imaging properties.

CHEMISTRY

Step by step

J. Am. Chem. Soc. **127**, 10462–10463 (2005)

The stepwise growth of a single polymer chain has been observed inside a 'nanoreactor'.

Hagan Bayley and Seong-Ho Shin of the University of Oxford, UK, studied how monomers link through the formation of bonds between sulphur atoms. The growing polymer chain was anchored to α -haemolysin, a bacterial protein that acts as a nanoreactor by shepherding the units into place. As the polymer grew, the conductance of the protein decreased, allowing the researchers to measure the lifetime of each intermediate over ten extension steps.

The authors suggest that the same technique could be used to study the kinetics of other polymerization reactions.

JOURNAL CLUB

Nicholas Spitzer

University of California, San Diego

The co-director of the Kavli Institute for Brain and Mind likes novel work on neural wiring.

Like many scientists, I'm drawn to big questions. And in neuroscience, one of the biggest is known as the wiring problem. This asks how the nervous system is wired up during embryonic development. With 100 billion neurons each making 10,000 synapses, the complexity

of the process is immense.

A consensus on what drives nerves to their targets had slowly emerged from decades of work. Then research by Gartz Hanson and Lynn Landmesser at Case Western Reserve University (*Neuron* **43**, 687–701; 2004) upset the apple cart. Previous studies suggested that electrical signalling played no part in the wiring process, but these researchers find that spontaneous electrical activity in chick embryos is necessary to guide the projections of motor neurons (axons) to muscles.

Everybody believed that axon pathfinding was driven by signals intrinsic to the cell, defined by regulatory proteins known as transcription factors, in concert with certain external guidance molecules. The evidence is strong — altering the expression of either component leads to altered wiring.

Some of my own research had hinted at a role for electrical signalling in pathfinding, but Hanson and Landmesser discover a link that makes their result really powerful. They show that

the pattern of electrical activity affects the expression of guidance molecules, although they have yet to demonstrate a connection with transcription factors. This work relies on an intimate knowledge of the patterns of activity in the embryo, which the researchers went to some pains to collect.

It is interesting that it's the pattern of activity rather than the total amount that's important, and I am delighted to reorient my thinking on the topic. Now, what drives the spontaneous activity?

NEWS

Senator boosts chances of stem-cell reform

WASHINGTON DC

Prospects for US stem-cell research brightened considerably last week when a key Republican senator backed the idea of loosening funding restrictions on the work.

In a speech on the floor of the Senate on 29 July, majority leader Bill Frist (Republican, Tennessee) endorsed efforts to increase federal funding for research on newly derived human embryonic stem-cell lines. He said that President George Bush's policy of limiting the use of federal funds to a handful of lines derived before 9 August 2001 needed changing.

"I believe the president's policy should be modified," Frist said. "We should expand federal funding and current guidelines governing stem-cell research, carefully and thoughtfully staying within ethical bounds."

Frist's announcement makes it much more likely that the Senate will pass legislation similar to that already passed by the House of Representatives, which voted to loosen funding restrictions in May (see *Nature* 435, 544–545; 2005). Research advocates even say that Frist's speech might make it possible for the Senate to later override a promised presidential veto of the bill — although the return of the bill to the House is unlikely to gather similar levels of support.

And whether or not the bill passes this year, they say, Frist's speech marks a turning point in the US debate on stem-cell research, because of his highly visible role in the Republican party and the Senate. "The ramifications of this are huge," says Kevin Wilson, director of public policy at the American Society for Cell Biology.

Frist had said recently that he was opposed to modifying the president's policy, and his change of mind was a surprise to many people involved in the stem-cell debate. But *Nature* has learned that Frist consulted with at least two scientists just days before his speech.

On 27 July, Frist spoke to Irving Weissman, a stem-cell pioneer at Stanford University and an outspoken critic of the president's policy. Weissman told him that the stem-cell lines currently approved for research cannot be used for therapeutic trials in people because they are probably contaminated with mouse viruses. He also explained that US companies are likely to need licences to develop therapies using the best techniques in the field, which have been pioneered by South Korean researchers.

"I told him that prohibiting a line of research has consequences, not just from a scientific perspective, but also from both economic and health perspectives," Weissman says.

The fact that Frist's speech placed strong



Bill Frist heads for the Senate to announce his support for changes to rules on stem-cell research.

emphasis on the development of potential treatments is encouraging, Weissman adds. "I knew something was going to happen, but I

More falling foam puts shuttle programme in serious doubt

After an embarrassingly large chunk of foam fell off the external fuel tank of the space shuttle *Discovery* during its 26 July launch, NASA has suspended further shuttle flights until the problem is solved. But as the agency has already spent two years and well over \$1 billion trying to make the shuttle safe, critics say there will be no quick solution.

A similar piece of foam fell off Columbia's fuel tank during take-off in January 2003. The hole it punched in the shuttle's wing caused the craft to burn up on re-entry, killing all seven astronauts inside. At the insistence of the Columbia Accident Investigation Board (CAIB), NASA has poured resources into ensuring the safety of future missions, in particular to secure the insulating foam that prevents ice from building up on the fuel tank.

Although the foam that came off *Discovery*'s tank last week didn't hit the craft, the size of the chunk, which weighed about 400 grams, shows that despite all the effort the problem is as big as ever.

Agency administrator Michael Griffin says it will be fixed "in short order", and has put together a 'tiger team' to look for answers. But many engineers question what NASA can do that it hasn't tried already. "Unless there is a significant redesign, there will always be a safety issue with this foam," says Henry McDonald, former head of the NASA Ames Research Center in California, and now at the University of Tennessee at Chattanooga.

Developing new foam could take at least a year, he says, with redesigns to the tank taking even longer. As the ageing shuttle fleet is due to be decommissioned in 2010,

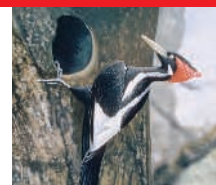
McDonald argues that NASA should now cut its losses and stop shuttle flights for good.

Doug Osheroff, a physicist at Stanford University in California, and a member of the CAIB, agrees that small tweaks won't help much, but major changes could take years. "We clearly don't understand all the mechanisms for foam shedding," he says.

The best way for NASA to quickly reduce the risk to the shuttle crew is to fly with fewer people, Osheroff says. "There's no reason to go up with seven astronauts."

As *Nature* went to press, *Discovery*'s crew was preparing to make emergency repairs, unrelated to the foam incident, to the craft's underside. For the latest news on the shuttle's progress, see www.nature.com/news/specials/returntoflight.

Mark Peplow



IVORY-BILLED WOODPECKER RAPS ON
Sound tapes convince critics that the extinct bird has survived
www.nature.com/news

G. M. SUTTON/CORNELL LAB ORNITHOL./NEWS.COM



C. SOMODEVILLA/GETTY IMAGES

Bone cells linked to creation of fresh eggs in mammals

A claim that stem cells in bone marrow and blood can restock mammalian ovaries with eggs is raising hackles among reproductive biologists. If true, the finding opens up avenues for delaying the menopause and preserving fertility in female chemotherapy patients. It also raises issues for women who have had bone-marrow transplants, by implying that subsequent children could be the genetic offspring of the donor.

Supporters of the work, which is headed by Jonathan Tilly at Harvard Medical School, have hailed it as a compelling challenge to the standard view of how ovaries work. "I see amazing implications coming from this work," says Kutluk Oktay, a physician at New York Presbyterian Hospital who pioneered ovarian transplants in women. But critics are dismayed that Tilly is already discussing the implications for women when, they say, he has yet to prove his case in mice.

Tilly first caused a stir in 2004, when his team published a paper suggesting that adult mouse ovaries can produce new eggs (J. Johnson *et al. Nature* **428**, 145–150; 2004). The work countered the view that female mammals are born with a store of eggs, and that when the store runs low, the ovary shuts down and menopause ensues. Biologists are still debating the claim. "It has never been reproduced as far as I am aware," says Allan Spradling, a developmental biologist at the Carnegie Institution in Baltimore.

Now, Tilly has proposed something even

more radical. His team found that bone-marrow stem cells in both mice and women express genes typical of germ cells. In mice, these genes cycle in unison with the same markers in the ovary (J. Johnson *et al. Cell* **122**, 303–315; 2005). Tilly proposes that these stem cells can travel to the ovaries, and that ovaries might signal to bone marrow, via an unidentified factor, for new stocks of eggs. "That factor could be of immense value therapeutically," he says, for example in treating premature menopause (see page 606).

To test the idea, his team transplanted bone marrow or blood cells to mice that

were either genetically sterile, or which had been given doses of chemotherapy that should destroy their eggs. Within two months of the bone-marrow transplants,

the researchers say, the mice regenerated hundreds of follicles — eggs encased by surrounding cells — at various stages of development, that persisted for at least a year. "That was amazing," says Tilly. And just 30 hours after the blood transfusions, several new eggs were visible.

Such rapid restocking leaves other researchers incredulous. Alan Trounson, a stem-cell researcher at Monash University in Melbourne, Australia, says that the quick appearance of eggs is unexpected and surprising.

Critics also doubt whether the new eggs come from the transplants. Tilly labelled his blood-transfusion cells with a protein that glows green, then showed the dye was present in the eggs. But Spradling says the eggs could have absorbed the dye from the blood.

Tilly's supporters argue that such scepticism is an understandable reaction to a radical idea. "The paper is an outstanding challenge to a dogma," says Oktay, adding that the idea is consistent with his finding that ovarian transplant patients seem to ovulate for longer than expected. "I don't think any revolution could be bloodless."

To prove the case, everyone agrees that Tilly must produce baby mice from eggs that come from bone-marrow transplants or blood transfusions. "We have tons of experiments under way to address this," he told *Nature*. "Should we show that, it's case closed."

Claire Ainsworth

"The paper is an outstanding challenge to a dogma."

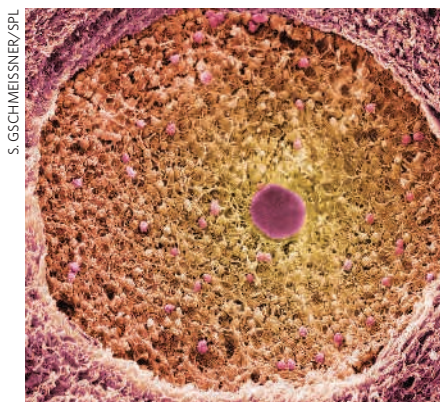
was surprised by how far he went," he says.

But Frist's change of heart still leaves it unclear whether he will give his full support to the core measure of the proposed bill, which would allow researchers to use federal funds to work on any embryonic cell lines. Frist said that he supports research into methods of producing human embryonic stem-cell lines that don't involve the use of a viable embryo.

In recent weeks, senators have proposed a flurry of bills supporting such methods — none of which has yet been shown to work (see *Nature* **436**, 309; 2005). *Nature* has been told that two senators, Kay Bailey Hutchison (Republican, Texas) and Norm Coleman (Republican, Minnesota), have also suggested 'compromise' bills. One would allow funding for research on cell lines created since the president's policy was announced until now. The other would allow researchers to use only 'spare' embryos created for *in vitro* fertilization currently existing at fertility clinics.

These bills could still pull Senate support away from the core measure — passed by the House and favoured by most scientists. As majority leader, Frist gets to decide how and when to put each bill to the vote when the Senate reconvenes next month after a long recess.

Erika Check



S. GSCHMEISSNER/SPL

Follicles may regenerate and produce new eggs, thanks to stem cells in bone marrow and blood.

ON THE RECORD

“Some never make a pregnancy, some die before term, some die of disease, some live but are large. And some are just fine — those are the real mysteries.”

Reproductive biologist Mark Westhusin describes the lottery of cloning. He brought the first cloned cat and deer into the world. The first cloned dog is reported on page 641.

“This hair sample could not have come from the hairy bipedal animal that was reportedly sighted in Teslin last week.”

Geneticist David Coltman dashes hopes of proving Bigfoot's existence, after mitochondrial DNA from hair found near a suspected sighting produced a 100% match with bison sequences.

SCORECARD

**Room to think**

Between 2001 and 2003, the floor space dedicated to scientific research at US colleges and universities rose 11% to 16 million square metres, making it a real growth area.

**Remains of the day**

Fossil hunters could soon get a chance to seek signs of life on Mars. NUGGET, an instrument proposed by NASA, would use neutron beams to peer inside rocks and draw three-dimensional images of any remains within.

NUMBER CRUNCH

NASA TV is taking over the world...

50,000 The peak number of simultaneous webcast streams sent out by NASA for the Mars Rover landings in January 2004.

118,000 The peak number of streams sent out for the Deep Impact mission on 4 July 2005.

433,000 The peak number of streams sent out for the space shuttle launch on 26 July 2005.

Source: NASA

Shadow hangs over research into Japan's bomb victims

On 6 and 9 August 1945, nuclear bombs dropped by the United States flattened the Japanese cities of Hiroshima and Nagasaki. The explosions killed 120,000 people; at least twice as many have since died from the effects of being exposed to radiation.

The attacks changed the political landscape for ever, but they have also made a major impact on science. The Radiation Effects Research Foundation (RERF) laboratories in Hiroshima and Nagasaki have tracked more than 80,000 survivors of the bombings, and the knowledge gained now forms the basis of our understanding of how radiation causes cancer, as well as the radiation protection standards used in hospitals and nuclear facilities. Yet 60 years on, uncertainties over US funding, and the foundation's ageing facilities, are causing fears about the future of a project that is still far from the end of its scientific life.

Over the past six decades, researchers at RERF have collected vast amounts of data from surveys, health examinations and biological samples, providing a resource for epidemiologists around the world. The survivors were a normal slice of the Japanese population. And unlike subsequent disasters such as Chernobyl, where some of the exposure came later through environmental contamination, almost all the radiation was delivered when the bomb detonated, making it easier to estimate each person's dose.

“We are talking about the biggest epidemiological study ever conducted,” says Yuri Dubrova, an expert on the effects of radiation on the human germline at the University of Leicester, UK. “The data from RERF produced information that is still the gold standard in terms of cancer studies. This is ongoing and astoundingly important research.”

Charles Land of the US National Cancer Institute in Bethesda, Maryland, says that with some 45% of survivors still alive, the most useful information may be yet to come. “There's still a lot to learn,” he says. “As the exposed population ages, there is going to be a lot more information coming from the survivors.”

Studies over the past decade or so are also starting to suggest that exposure to radiation increases the risk of many diseases besides

cancer, and that there may be no safe dose for treatments such as radiotherapy (see ‘Fallout beyond cancer’, below).

But despite the project's value, many experts are worried about its future. The foundation's budget has been around ¥3.8 billion (US\$33.8 million) for several years. Until 1996, half of that came from the United States, but its contribution has dropped to 38%.

Then last year, the US Department of Energy (DoE) suggested cutting its annual contribu-

“The foundation's facilities in Hiroshima, housed in prefabricated huts built in 1949, are run down. Roofs leak and leaves blow into labs.”

tion. Full funding was eventually delivered four months late and is guaranteed for another two years, but some are edgy that a similar situation could arise again. “Let's hope the DoE doesn't try any more of the tricks it did last year,” says Mark Little, an epi-

demologist at Imperial College London. “It's vital they keep going for the foreseeable future.”

Little has worked with data from RERF for the past 15 years, and was lead author of a letter sent to *The Lancet* last year in protest at the proposed budget cuts (M. P. Little *et al. Lancet* 364, 557–558; 2004). He says that the foundation's facilities in Hiroshima, housed in prefabricated huts built in 1949, are run down. Roofs leak and leaves blow into labs. There are also staffing problems. “There have been significant departures of epidemiologists,” says Little. “They are struggling

Fallout beyond cancer

Over the past decade, it has become clear that those who witnessed the Hiroshima and Nagasaki bombings do not just face an increased cancer risk. They also have higher rates of high blood pressure, stroke, and liver and kidney problems, among others.

When first identified in the early 1990s, some doubted that these effects were real, as there was no obvious way for radiation to cause such diseases. But more data have strengthened the evidence and the effects are now widely accepted. Around 4,500 survivors have died of heart disease since the bombs, for example, and 100 of these are associated with the radiation.

“We've been discovering what in the past would have been unthinkable as damage from radiation exposure,” says Kazuo Neriishi, who



HULTON ARCHIVE/GETTY IMAGES

Out of the ashes: survivors of Hiroshima and Nagasaki form 'the biggest epidemiological study ever'.

to do the analysis on the latest samples.”

Dale Preston, an epidemiologist formerly at RERF, shares Little's concerns. “Over the past few years people have been less keen to work in this area,” he says. “But there's nothing that can do the kind of work RERF can do.”

Kazunori Kodama, chief scientist at RERF,

says that work is continuing for the time being. “The place is holding out right now,” he says. “But I'm still worried about what will happen after 2007.”

Tom Simonite

Additional reporting by Jim Giles in London, and Ichiko Fuyuno and David Cyranoski in Tokyo

studies cataract risks at the Radiation Effects Research Foundation in Hiroshima.

Some researchers are now asking what the bomb data might reveal about the risk to radiotherapy patients. Comparing the groups is complex, as the survivors received a dose across their bodies. In radiotherapy, the radiation is focused on the tumour. But some epidemiologists believe that there is enough evidence to warrant further studies.

Sarah Darby, an epidemiologist at the University of Oxford, UK, says that the relationship between radiation dose and heart disease for the survivors follows a smooth curve. This shows that there may be no cut-off point, with even low doses affecting disease rates.

In a study published online last month, Darby and her colleagues show that British women

who had received radiotherapy for cancer in their left breast, the side on which the heart is located, were up to 50% more likely to die of heart disease than those whose right breast was treated (S. C. Darby *et al. Lancet Oncol.* **6**, 557–565; 2005).

Darby says radiotherapy doses have since been reduced in the United Kingdom, but she believes the bomb data show that patients may still be at risk. She is trying to investigate this through further studies of radiotherapy patients.

Researchers say they are some way from being able to confirm the mechanism behind the increased heart-disease rates. Some suspect that it may lie in a chain of physiological changes set off by the immune system's response to the initial radiation dose.

J.G.



US RICE CARRIES AN ARSENIC BURDEN

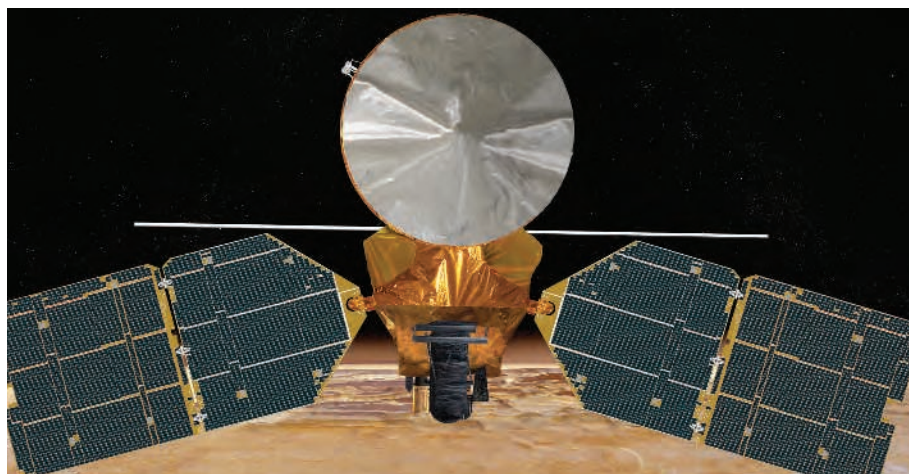
Legacy of cotton pesticides might be poisoning crops.
www.nature.com/news

Mars orbiter ready to scout for future landing sites as NASA looks ahead

With two healthy rovers still roaming the martian surface 19 months after they landed, NASA is set to take the next step in Mars exploration by sending its largest orbiter to the planet since the 1970s. The \$500-million Mars Reconnaissance Orbiter (MRO), due to launch from Cape Canaveral, Florida, on 10 August, may also signal a new period of stability for the agency's Mars programme. After recent budget scares and shifts in priorities, the programme looks to be on track to send more advanced rovers to the planet over the next decade.

After it arrives in March next year, the MRO will spend eight months settling into its final 320-kilometre orbit. From there it will point six instruments at the planet's surface, including a spectrometer for identifying minerals, a sounding radar for locating subsurface water (see *Nature* 435, 266–267; 2005), and a camera sharp enough to resolve objects the size of a kitchen table. The camera will increase high-resolution coverage of the planet from 2% to 20%, and allow Mars planners to scout for future landing sites.

The MRO will be followed in 2007 by the Phoenix lander, designed to search for water and organic molecules around the planet's north pole. Then in 2009 the most ambitious mission in the queue, the mobile Mars Science Laboratory (MSL), will launch. With a price tag of at least \$1 billion, the MSL will travel a kilometre or more from its landing site carrying a suite of sophisticated instruments for sniffing out chemical evidence of life.



NASA/JPL

The Mars Reconnaissance Orbiter will analyse conditions on the red planet in unparalleled detail.

This spring, the MSL was in serious danger of having its launch date slip two years or more as NASA struggled to contain rising costs. But the situation has improved, and the mission remains targeted for 2009, says Mars programme director Douglas McCuistion. This reprieve is partly due to a decision to defer missions that will pave the way for astronauts to visit Mars around 2030. McCuistion says that launching two MSLs, which science advisory groups have recommended partly as insurance against a launch failure, is “not impossible” but depends on future NASA funding levels.

The money crunch has also eased with NASA's decision to cancel the 2009 Mars

Telecommunications Orbiter, which would have acted as a high-speed data relay for other Mars spacecraft. McCuistion says the MRO and other spacecraft already orbiting the planet will serve as data relays instead.

Bringing samples home

The new Mars plan retains a sample-return mission as a goal, but it has an unspecified launch time. University of Colorado planetary scientist Bruce Jakosky, who until recently chaired NASA's outside advisory group for Mars exploration, says there are “legitimate differences of opinion” about where sample return should fit in the mission queue. Most scientists still consider bringing Mars samples back to Earth a priority, says Arizona State University planetary scientist Ron Greeley, as the rocks must be dated isotopically in a lab to sort out the planet's complicated geological history.

But bringing samples home will be difficult and expensive. And new, more stringent recommendations from the National Academy of Sciences for sterilizing Mars-bound spacecraft, to protect against biological contamination of the planet, could drive up the cost even more.

Yet NASA's long-term plan to send astronauts to Mars could ultimately work in favour of sample return, as engineers will need to understand how surface materials behave and what risks they might pose. Says Greeley: “I'd be surprised if we didn't have sample return as a precursor to humans for safety reasons.”

Tony Reichhardt

NASA'S FUTURE MARS MISSIONS

HST/NASA



10 August 2005: Mars Reconnaissance Orbiter

Will make high-resolution measurements of the surface from orbit after its arrival in March 2006.

Late 2007: Phoenix lander

Will search for water and organic molecules in the high-northern latitudes of Mars.

Late 2009: Mars Science Laboratory

A science lab on a mobile rover. Will look for signs of life, as well as studying local geology and planetary processes relevant to prospects for life.

2010+: Future missions

Plans include at least one Mars Scout mission and possible sample-return missions, although there are no set dates. Work is planned to develop miniaturized surface-science instruments and the ability to drill 100 metres below the surface. Still aiming for a manned mission around 2030.

Drugs could head off a flu pandemic — but only if we respond fast enough

If a strain of avian influenza emerges that can spread easily from person to person, could rapid deployment of antiviral drugs stop a local outbreak from becoming a global disaster? Yes, conclude the most detailed modelling studies yet of an emerging pandemic — if the world can muster its scientific and logistical efforts quickly enough.

The two independent studies were carried out by an international team led by mathematical biologist Neil Ferguson of Imperial College London¹, and by a group led by Ira Longini, a biostatistician at Emory University in Atlanta, Georgia². They reach markedly different conclusions about how easy it would be to contain an emerging pandemic. But both agree that it would be possible — if the virus was detected quickly, if it did not spread too fast, if sufficient antivirals were deployed quickly and massively around the outbreak's epicentre, and if strict quarantine and other measures were also used.

That in itself is something of a breakthrough: all previous pandemics have swept unchecked across the planet. "What is striking is that these models, although quite different, both conclude that a flu pandemic could be contained at source," says Jeremy Berg, director of the US National Institute of General Medical Sciences in Bethesda, Maryland.

Buying time

Treating sick patients with antivirals is not enough to stop a flu pandemic. You have to at least treat their contacts as well, because patients cough out virus before they fall ill, as do others who don't get ill at all.

But must you treat just those in contact with the patient, or everyone within, say, a 1-, 10- or 100-kilometre radius? How many drugs would that take? And how fast would they need to be delivered?

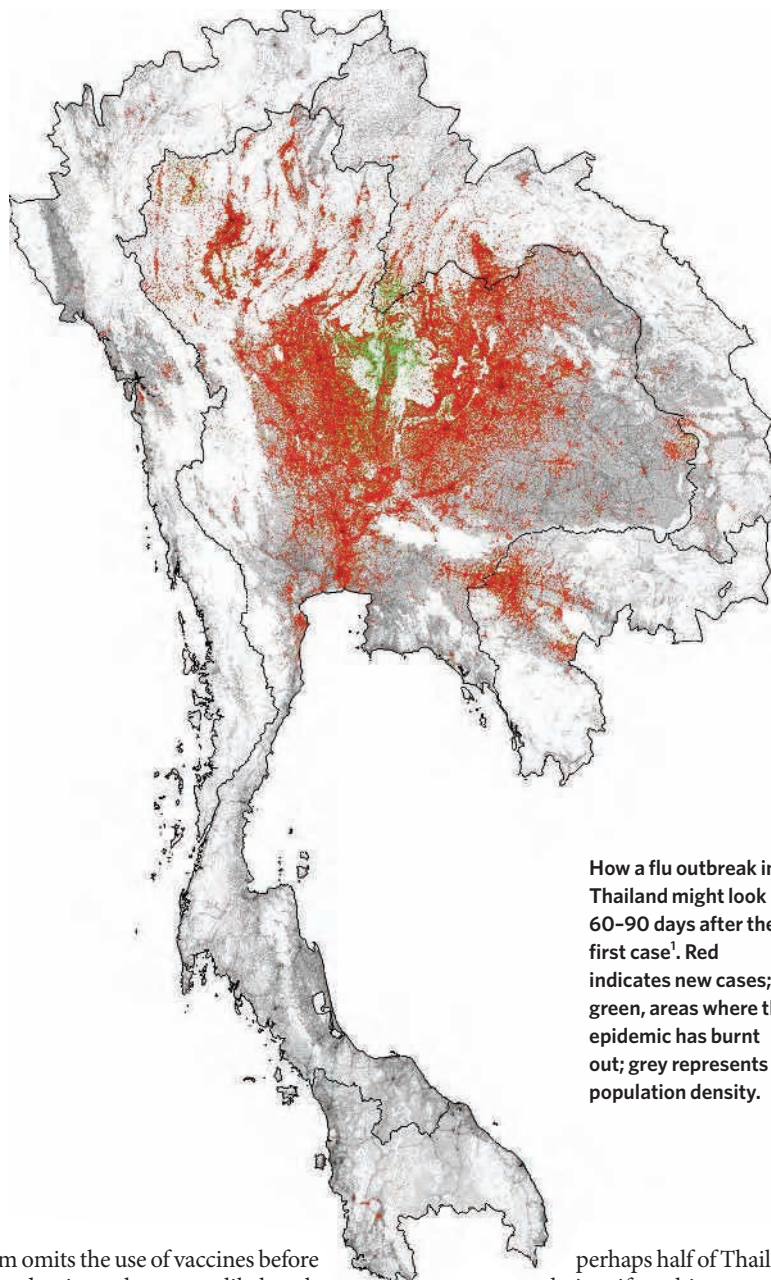
The studies, which are both based on data from Thailand, give different answers. Longini and his colleagues found that just 100,000 to 1 million drug courses, administered to those who fall ill and their social contacts, would give a high probability of success. They also predict that even a rapidly spreading virus could be contained by extra measures such as quarantine and pre-pandemic vaccination. "There is a lot of fatalism about avian flu," says Longini. "But we are saying it is perhaps not hopeless."

Ferguson's scenario is more pessimistic. His

team omits the use of vaccines before a pandemic, as they are unlikely to be available. It concludes that a slowly spreading pandemic might be stopped, but it would mean treating everyone in a 5-kilometre radius, involving some 2 million to 3 million drug courses and measures such as quarantine from the start. If the policy succeeded, only 200 people might get infected, compared with

perhaps half of Thailand's population if nothing was done. The model also concludes that a faster-spreading virus would be unstoppable, although it might be held back for a few weeks, buying precious time for a vaccine to be developed, something that takes 6–8 months.

One reason for the differences in the studies' results is that the groups calculated the initial



How a flu outbreak in Thailand might look 60–90 days after the first case¹. Red indicates new cases; green, areas where the epidemic has burnt out; grey represents population density.

NEIL FERGUSON



BIRD FLU MOVES TOWARDS EUROPE

Migratory birds may have caused outbreaks in Russia and Kazakhstan.

www.nature.com/news

rates of viral spread differently. Longini's group assumed it takes four days for an infected individual to be able to infect others, a figure used in previous models. But Ferguson reanalysed historical data and came up with a figure of just 2.6 days.

Longini's simulation also models 500,000 individuals laid out on a regular grid, whereas Ferguson's maps the population densities of all of Thailand's 85 million people, albeit in less detail. The larger scale makes it easier to take account of clusters arising outside the initial outbreak area.

Too slow

Both groups agree that, for a containment strategy to have any hope of working, it must be in place within a few weeks at most of the first people being infected with a virus capable of sustained human-to-human transmission.

If such a virus arose today, that is unlikely to happen. Surveillance systems in southeast Asia are poor; recent cases have taken weeks to detect and diagnose. Whereas Cambodia has typically reported cases to the World Health Organization (WHO) within about a week, Vietnam has often reported cases after several weeks, and in some cases months.

Marc Lipsitch, an epidemiologist at Harvard University, says the papers leave him concerned that too little is being done to plan containment strategies. "We are simply not moving fast enough," he says.

For example, the WHO currently has just 120,000 courses of antivirals in its stockpile, although it is in discussions to get more. "I think the take-home message is that the current stockpile is very unlikely to be adequate to stop anything," says Lipsitch.

What's needed, says Ben Schwartz of the National Immunization Program at the Centers for Disease Control and Prevention in Atlanta, are international agreements on how to investigate and report clusters; training and resources to strengthen surveillance; and measures to ensure that the WHO has enough antiviral drugs. The countries where a pandemic is most likely to emerge need detailed plans and drills, he adds.

The \$25 million spent by the United States last year in boosting surveillance in Asia is inadequate, says Schwartz. He points out that the country spent more than \$800 million on anthrax vaccines, "against a pathogen that has killed only a handful of Americans and whose bioterrorist potential is unproven". ■

Declan Butler

1. Ferguson, N. M. *et al.* *Nature* doi:10.1038/nature04017 (2005).
2. Longini, I. M. Jr *et al.* *Science* doi:10.1126/science.1115717 (2005).

US energy bill pushes research but fails to cut consumption

WASHINGTON DC

The US Congress slapped an energy bill, four years in the making, on President George W. Bush's desk last week.

The United States uses vastly more energy than any other country on the planet, and the bill was initially seen as a chance to set out a clear strategy for the country in terms of energy efficiency.

But in the end, critics say, the 1,700-page Energy Policy Act is more of a compromise than a strategy. It has been shorn of many of its controversial provisions, and won't do much to make the country's energy use more environmentally friendly, at least in the short term. But its various tax breaks and incentives may change the landscape of energy science.

In the past few months, sections of the bill protecting manufacturers of the water-contaminating petrol additive MTBE and opening the Arctic National Wildlife Refuge to oil drilling were scrapped so that Congress could finally pass it. The bill sets no emissions limits and does not change fuel-efficiency standards for cars. A proposal that 10% of US electricity should come from renewable sources by 2020 was also ditched.

In fact, energy efficiency and renewables take home just \$5 billion of the bill's \$14.5 billion in tax incentives, which are spread over ten years. The rest is

largely a list of benefits for traditional energy industries, including a \$1.5-billion scheme for research and development into drilling for oil and gas in the Gulf of Mexico.

The bill may also pave the way for a resurgence of the nuclear industry in the United States, which has not signed off a new nuclear-plant construction since 1973. Energy companies interested in ending that streak can now count on a tax credit and reimbursement of any losses associated with unforeseen regulations, although it is not yet clear whether the industry will bite.

Science seems to do well out of the bill, with more than \$30 billion assigned to various research and development programmes over three years. But these are really just a starting point for negotiations by the appropriations committee, which is widely expected to be more frugal.

One clear change, however, is the creation of an undersecretary for science in the energy department, a position that many physical scientists hope will increase the clout of research in the department's budget wrangles. "This provides a voice at the table where the crucial decisions are made," says Michael Lubell, head of public affairs at the American Physical Society. ■

Emma Marris



Guzzle on: the proposed US energy bill will not improve fuel efficiency in cars.

Possible planet prompts debate over definitions

The discovery in the Solar System of a body larger than Pluto has left astronomers racing to draw up rules to determine which objects qualify as planets.

The discovery of a possible tenth planet, currently named 2003 UB313, was announced by US astronomers on 29 July. This world of rock and ice is in orbit some 15 billion kilometres from the Sun. Planetary astronomer Mike Brown at the California Institute of Technology, one of the three-man team that identified the object, declared it to be a new planet and submitted a proposed name to the International Astronomical Union (IAU).

But the IAU, which oversees the naming of stars and asteroids, has no criteria for defining planets. A committee has been working on the issue for about a year and

had planned to publish its results next summer. The new discovery has now made the debate more “urgent”, says Iwan Williams, a committee member at Queen Mary, University of London.

Government dispute halts bird flu work at Chinese lab

A Chinese lab has stopped its work on H5N1 avian influenza following a disagreement with the country’s agriculture ministry.

The dispute began on 6 July, with a paper published online in *Nature* by the Joint Influenza Research Centre in Shantou. This reported a link between poultry in southern China and the recent flu outbreak among migratory birds at Qinghai Lake (see *Nature* 436, 191–192; 2005).

Chinese officials denied the link, then scolded the lab for not meeting safety standards and for not getting its work on the H5N1 virus approved by the government.

The researchers halted the project on 25 July, but argue that the lab meets World Health Organization standards. They also say that rules restricting work on H5N1 were not announced until after their paper had been sent to *Nature*. They have now applied for permission to continue their research.

Astronomy division calls for help to make cutbacks

Finding that it has more on its plate than it can afford, the US National Science Foundation’s astronomy division has asked a panel of outside scientists to help it identify some \$30 million in cuts between now and 2011.

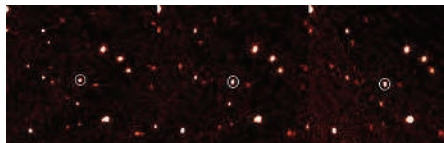
With an annual budget of less than \$200 million, the division is struggling to pay for ambitious new projects such as the Atacama Large Millimeter Array. The division’s director, Wayne Van Citters, admits that the \$30 million, which he calls “a target we may not be able to meet”, would be just a down-payment on a larger problem.

The review panel, chaired by astrophysicist Roger Blandford of Stanford University in California, will meet for the first time in October, with a final report expected next spring.

Invaders cause rash of problems for swimmers

Flatworm parasites living in a non-native Japanese snail may have caused more than 90 cases of a temporary skin rash in San

PALOMAR OBS



World view: time-lapse images of a possible tenth planet (circled), which is in orbit beyond Pluto.

K. RASKOFF/NOAA

Arctic survey plumbs the depths to find marine life



This delicate-looking jellyfish of the genus *Crossota* is just one of many other-worldly creatures recently plucked from the depths of the Arctic Ocean. A US-led team of oceanographers this summer explored the Canada Basin, an area isolated by high ridges

and a thick layer of floating ice. Remote-controlled vehicles brought back thousands of specimens, many from depths of 3,300 metres and below. The findings suggest, team leaders say, that creatures thrive at far higher densities than expected in the frigid Arctic waters.

San Francisco Bay in June. Experts say it may be the first time a US public-health problem has been linked to an invasive marine species.

The flatworm's identity is unknown, but its host, the Japanese bubble snail (*Haminoea japonica*), arrived in San Francisco Bay just five years ago — probably as larvae in ships' ballast water, says Andrew

Cohen of the San Francisco Estuary Institute. The snails may have enabled the flatworm to spread onto the beach where the swimmers' itch cases occurred.

Ships arriving in US ports from other countries are required to flush their ballast water before arrival. But the procedure leaves many organisms behind, says Cohen.

For instance, ballast water may have carried *Vibrio cholerae* to South America in 1991, triggering a cholera epidemic there.

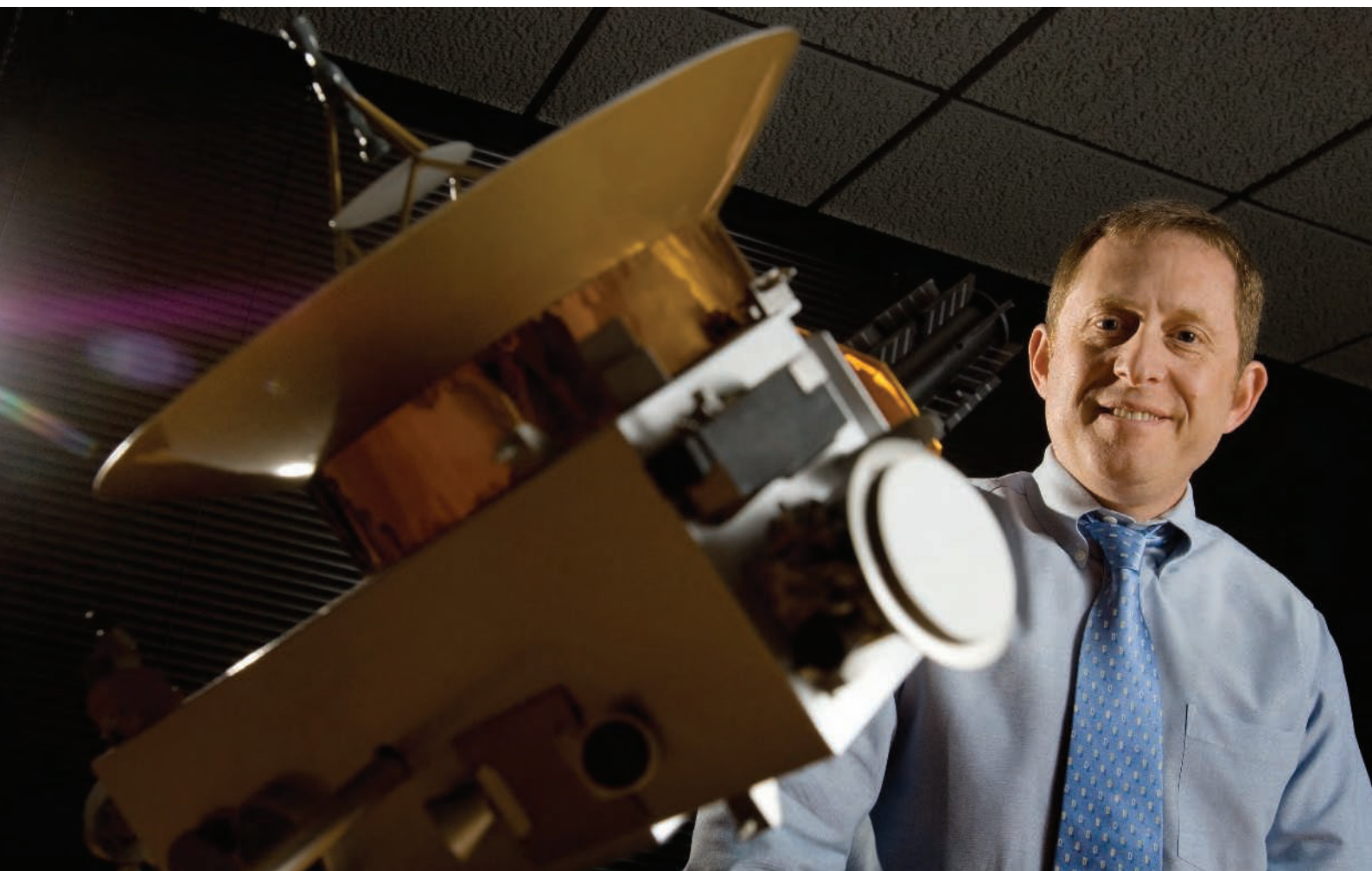
Critics deride climate pact as excuse for inaction

Climate-policy experts have criticized a six-nation pact that the Australian government claims is superior to the Kyoto Protocol on climate change.

On 28 July, Australia, China, India, South Korea, Japan and the United States signed the independent agreement, which they had worked out in secret. The resulting Asia-Pacific Partnership on Clean Development and Climate promotes the use of new technologies, such as renewable-energy systems and more efficient vehicles, to reduce the emission of greenhouse gases.

But the pact has earned scorn for not adopting specific emission-reduction targets. Australia and the United States are the only two developed nations not to ratify the Kyoto Protocol, and they are widely expected to use the new pact to deflect pressure to accept future versions of the protocol.

International talks about what to do when the Kyoto agreement expires in 2012 will begin in earnest in November.



A man with a mission

In 2015, Pluto will welcome its first visitor, a robot named New Horizons. **Amanda Haag** meets the planetary scientist who nursed the mission through its darkest days.

In the spaceflight business, delayed gratification is often part of the deal. The time between a launch and the scientific pay-off may be a decade or more. So for the man leading a mission that is about to begin its ten-year voyage to Pluto, the downtime of space travel promises one sure thing: more time for research.

Planetary scientist Alan Stern is principal investigator for NASA's New Horizons mission — slated for a January launch — to Pluto and the frosty worlds beyond. He has flown in fighter jets to conduct upper-atmosphere experiments, sent instruments to visit comets, and qualified as a mission specialist for the space shuttle. In between, he led the assembly of one of the world's most elite and creative groups of planetary scientists, based at the Southwest Research Institute (SwRI) in Boulder, Colorado.

Of all these accomplishments, Pluto is likely to be the most public legacy of this 47-year-old space junkie. For years, Stern was for the most part a lone voice arguing for a mission that looked as if it might never get off the ground. Various incarnations of a Pluto spacecraft had been funded and cancelled over a period of 12 years before Stern resurrected the notion in the form of New Horizons. But even then, a two-year funding battle ensued, with New

Horizons' future hanging by a budgetary thread. "It took more years to get this mission out of the Washington beltway than it does to cross the entire expanse of the Solar System," he says, only half-joking.

It didn't take that long for the space bug to bite the young Stern. Like many children in the 1960s, he grew up hugely inspired by the Apollo space programme. "Everybody had an astronaut helmet and a little silver Halloween suit," recalls Stern. As a teenager, he was the quintessential space nerd. "Capitalize the 'N,'" he jokes. But that fascination didn't always translate to scholastic accomplishment. After years in a strict boys' prep school in Dallas, Stern relished the freedom at the Austin campus of the University of Texas so much that he nearly flunked out and left. Six months later, after working as a dock hand and cutting chicken for a fast-food restaurant, he decided to give academia another shot.

Launched anew

With his trademark intensity, he soon acquired both double undergraduate and double master's degrees. Two years in the aerospace industry convinced him to look at academia as a career option, and in 1983 he landed work at the Laboratory for Atmospheric and Space Physics at the University of Colorado's Boulder

K. MOLONEY

campus. Stern recalls spending one winter in Alaska shooting a sounding rocket into the aurora borealis as a satellite flew through the lights. Soon, one of his primary jobs became helping to build a pair of ultraviolet spectrometers that would study Comet Halley from aboard the space shuttle Challenger. As the project unfolded, Stern gravitated towards the science rather than the engineering. "It really swept me off my feet," he tells me from his downtown Boulder office, with its stunning view of the jagged Flatiron Mountains. "I decided I wanted to be a space scientist as a result."

But his hopes for the project were dashed when Challenger, carrying the Comet Halley experiment, blew up just after launch on 28 January 1986, bringing the US space programme to a standstill. "It looked like everything we were doing in my work would be put off for years," recalls Stern. So within a month, he decided to go back to graduate school. "If there was ever going to be a time to become a scientist, this was probably it."

Building on the vision

Less than two years after completing his doctorate at the University of Colorado — which he did in just three years — Stern was hired to forge a planetary-science group at the SwRI's offices in San Antonio, Texas. The institute was already well known for its strong space-physics research group, but it wasn't on the map for planetary science. "Most places you go, you're going to wait in line for 10 or 15 years to get your shot to lead," says Stern, "but at Southwest, there was nobody in line in front of me in planetary science. It was pretty much mine to make it — or not — which was simultaneously thrilling and terrifying."

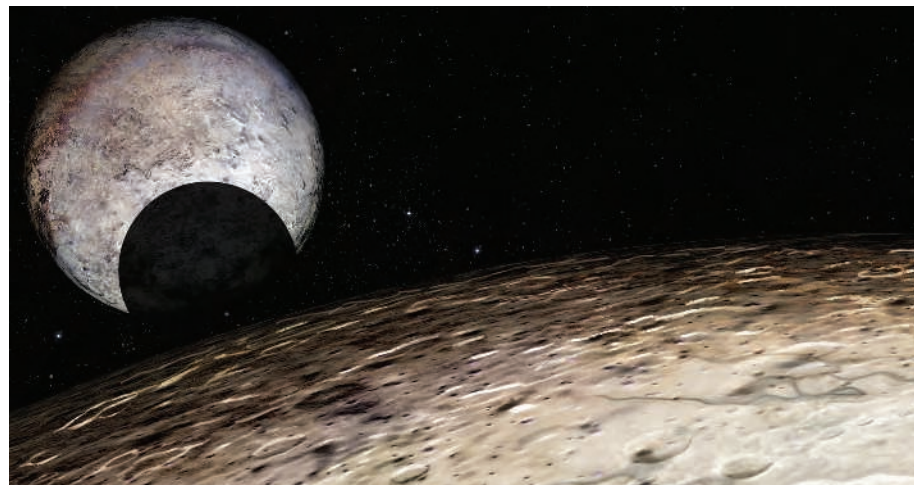
Although Stern acquired numerous research grants and won instrument proposals for space missions, he found San Antonio lacking key ingredients: local research universities and a community of planetary-science experts. Few recruits were interested in coming to the city. Then, at a 1993 banquet, Stern found himself seated next to three talented planetary scientists who had turned down offers to join the SwRI. Once the wine started to flow, they told him: "If you guys were in Boulder, or some place with a big research university, we'd be at Southwest." "That," says Stern, "was an epiphany."

Stern became convinced that in order to build the landmark planetary-research group he envisaged, he would have to take a bold step. So he and his boss, magnetospheric scientist James Burch, went straight to SwRI president Martin Golland, who led the institute for 39 years until his death in 1997. Golland threw Stern out of his office twice before finally relenting. Eleven years later, the Boulder office has grown from two scientists to more than 40, with a deep reach in both planetary science and solar physics.

To expand the institute, Stern drew hand-picked scientists from across the country. He lured Dan Durda, an asteroid expert, with the

promise of conducting airborne astronomy in fighter jets. Stern and Durda have since perfected a high-performance telescope to study asteroids near the Sun from a vantage point 17,000 metres away. Since joining the SwRI at Boulder, Durda's research has branched out into areas that he says he would never have imagined possible. A cave diver in his spare time, Durda is now developing technologies for exploring under the ice of Jupiter's moon Europa through a partnership at the SwRI. For practice drills, Durda drops his robotic vehicle into deep sinkholes in Mexico. "There is no limit to what you can accomplish here," he says.

Stern designed the department with characteristic calculation. One of his aspirations was to build an interdisciplinary department — funded by the soft money of research grants — that



Out of the shadows: Alan Stern and his team aim to cast light on Pluto and moon Charon, simulated here.

circumvented some of the classic problems of universities. Most importantly, the group is designed to prevent turf wars. By not allowing any given research group to sit together in one area, "there's no territory to claim, and people have to get up and walk past each other all the time," says Stern. The result is a place "that's really crackling with intellectual energy," says David Grinspoon, an SwRI scientist who studies the evolution of Earth-like planets. "You can't help but walk down the hall and before you know it get snared into some interesting project."

A wild success

Soft-money jobs are often viewed as transient positions. But the SwRI at Boulder hasn't lost a single scientist to recruitment, although many of its researchers have been offered positions at competitive institutes. "It's just been a success story wildly beyond what I ever thought would be possible," says Michael Shull, a University of Colorado astrophysicist and one of Stern's PhD advisers. "I think people probably saw it as a stopping point for a couple of years and then on to another job, but they've all ended up staying."

When he's not polishing his departmental efforts, Stern is busy working on comets and other objects in the icy realm beyond Pluto. This distant region some 5 billion to

8 billion kilometres from the Sun, known as the Kuiper belt, turns out to have swarms of frosty objects orbiting along with Pluto and its sister moon, Charon. "It's kind of the Solar System's attic, with all kinds of things stuffed away up there," says Stern, "like a huge number of small and moderate-sized 'planets' that no one expected even 20 years ago."

Stern is also working on NASA's Lunar Reconnaissance Orbiter, set to reach the Moon at the end of 2008 or in early 2009. He is principal investigator for the ultraviolet spectrometer on the orbiter, as well as for the European Space Agency's Rosetta mission, which will orbit its comet destination beginning in 2014. And if all goes well, the New Horizons mission will fly by Pluto and Charon in 2015. "I'm looking to have one heck of a

bang-up decade in the teens," he says.

However frenetic the pace, Stern also manages to be a husband and father of three. The first day we meet, he is rushing out the door to take his oldest daughter, Sarah, shopping for her first car. On weekends, he spends time individually with each of his children. His wife Carole says they've had to be creative to adapt to Stern's exhausting travel schedule, which often means he makes 40 to 50 trips a year.

So what does Stern hope will turn up when New Horizons flies by Pluto? At a conference in 1993, during an early incarnation of Pluto mission studies, scientists placed their predictions about what mysteries Pluto might hold into sealed envelopes. They'll be opened when a spacecraft makes it to the planet. "You're not supposed to tell what your prediction is," says Stern. Breaking into his characteristic grin, he adds: "But I'm going to tell you mine."

I lean forward, full of suspense. "I wrote," he says, "that when we get to Pluto — and this is my entire prediction in two words — we would find 'something wonderful'. We've never been to any two places that are the same in the Solar System. So we always learn something completely wonderful."

Amanda Haag is a freelance writer in Boulder, Colorado.

M. GARLUICK/SPL



The valley of ghosts

While other Asian tigers are roaring ahead in biotechnology, Malaysia's BioValley is going nowhere fast. **David Cyranoski** asks what went wrong.

Asking Malaysian researchers what happened to their country's flagship science project, known as the BioValley, is a confusing experience. Some claim it is still under development. Others say it never existed. Many are simply unwilling to talk about it.

But this was always a difficult project to pin down. Launched in May 2003, the BioValley was one of the final initiatives of Malaysia's strongman prime minister, Mahathir bin Mohamad, who stepped down from power a few months later. Incorporating three new research institutes and costing some US\$160 million, the BioValley was meant to attract biotech companies to a centralized hub that would offer cheap rent, good telecommunications infrastructure and access to the country's lush biodiversity — a potential source of new drugs and other useful products.

But even after its launch, it was hard to obtain concrete details about the BioValley. Aside from the plans drawn up by famed Japanese architect Kisho Kurokawa, the project was shrouded in mystery. By now, the 80-hectare campus in Dengkil, south of Kuala Lumpur, should have been nearing completion. Instead, the site lies empty. And official documents reveal that, earlier this year, the BioValley quietly morphed into the BioNexus, a much less ambitious scheme comprising just one new institute in Dengkil, and two other 'centres of excellence' built around existing labs elsewhere.

All this is in marked contrast to developments in neighbouring Singapore, the city-state that nestles at the tip of peninsular Malaysia. There, a formidable biomedical research hub, the Biopolis, positively bustles with activity.

Problematic past

On the face of it, the disparity is puzzling. Singapore and Malaysia have much in common — their populations have a similar ethnic mix, both have governments with an authoritarian streak, and both see biotechnology as a springboard for future economic growth. Malaysia, in particular, wants to decrease its heavy reliance on the electronics industry and the production of palm oil.

But while Singapore has recognized that scientific success means aggressively recruiting top talent regardless of nationality, race or creed, Malaysia's biotech push has been hampered by a legacy of ethnic strife, its hands tied by an educational policy designed to favour its ethnic Malay majority.

The BioValley is just the most conspicuous feature in a landscape of failed effort. Elsewhere, flashy new labs remain largely unused, some of them led by people without proper scientific credentials. And in a culture in which criticism of authority is taboo, these problems don't look remotely near resolution. One senior political figure (who, like most of the people interviewed for this article, did not want his name mentioned) complains that the

BioValley "was all about fancy buildings and real-estate development".

Mahathir and his acolytes seemed to assume that researchers would come pouring into shiny new centres bearing the label 'biotechnology'. It was a naive view, suggest foreign observers familiar with the Malaysian scientific scene. "With no history in biotechnology, and little industrial presence, the risk is very high," says Keiichi Kiyota, president of the Tokyo-based Nimura Genetics Solutions, one of very few foreign companies with research activities in Malaysia. "The greatest problem is the lack of manpower," he adds.

Given this dearth of talent, Malaysian science can ill afford the brain drain that sees many young scientists, particularly those from the nation's Chinese and Indian minorities, leave the country. It's easy to see why, given that the dice are loaded against them. "The 'Malays first' policy holds them back," says biochemist Barry Halliwell, who heads the National University of Singapore's graduate school. "It does Singapore a good favour, as many come here." Last year, for instance, 128 students with straight A grades were denied access to medical school in Malaysia, while less qualified candidates were accepted. The excluded students were all non-Malay.

The 'Malays first' policy has its origins in the race riots of 1969, which were sparked by the Malay majority's resentment of the Malaysian Chinese community's economic successes. Given the bitter memories of this conflict,

some researchers back the policy of granting privileged opportunities to Malays. "Otherwise people would become second-class citizens in their own country and you'd have a time bomb on your hands," says Salleh Mohammed Nor, former director of the Forest Research Institute of Malaysia in Kepong, near Kuala Lumpur, and now president of the Malaysian Nature Society.

In the early 1970s, the government made a concerted effort to promote the interests of the Malay majority. In 1975, for example, the Malay language — Bahasa Malaysia — replaced English as the standard language of education. But critics say that this policy has damaged Malaysia's education system by failing to reward merit. "All vice-chancellors are appointed by the government without any kind of search committee," says one former University of Malaya researcher. "It's all favouritism."

Empty labs

Even when new labs have been built, they've failed to make much impact. The Technology Park Malaysia near Kuala Lumpur, for instance, hosts a government-sponsored institute that was supposed to act as a magnet for biotech companies. When *Nature* visited the two-year-old facility in late June, its high-performance liquid chromatography and mass spectrometry instruments lay idle — and only two research staff were present, huddled by a computer. Malaysia has unemployed graduates, but many don't have the requisite skills, including English ability, says an administrator at the park. "Good people go overseas," he adds.

This failure to embrace the international language of science is symptomatic of a general detachment of Malaysia's research system from the world scene. For most Malaysian researchers, publications in international peer-reviewed journals do not seem to be a priority. "People here don't seem to publish much, apart from in workshop and conference proceedings," says one visiting ecologist.

The country has also attracted few foreign researchers. Pay is low and there are few post-doctoral students to work with unless you bring your own. "There is nobody here who really understands what I am doing apart from my students," says a foreign researcher who is in Malaysia for family reasons. "People in my department are perpetually putting obstacles in my way."

Again, the contrast with Singapore is stark. Researchers there have high pay and high status, and the government has cast its net wide to bring in top scientific talent. Of the 35 principal investigators at the Institute of Molecular and Cell Biology, the country's premier research centre, only one is Singaporean. "If people have brains, I'll borrow them," declares Philip Yeo, who chairs A*STAR, the country's main science funding agency.

In theory, Malaysia's leaders recognize the need to emulate Singapore's hiring policies. In 1995, for instance, Mahathir initiated a five-



Industrious: but many Malaysian labs are woefully understaffed.

year plan to recruit 5,000 foreign researchers a year. But the scheme attracted just 94 scientists, and 24 of them were returning Malaysians. By 2004, only one of these researchers remained in the country.

This pattern of setting and then failing to meet grandiose targets was common in the Mahathir era. So it should perhaps come as no surprise that the BioValley never made it off the drawing board. Its humbler successor — the BioNexus — is based around existing labs specializing in agricultural biotechnology, genomics and molecular biology. The single new centre will focus on pharmaceuticals and nutraceuticals.

The BioNexus is part of the national biotechnology policy that was unveiled in April this year, which is supposed to remedy previous failings. A new organization, the Malaysian Biotechnology Corporation, is chaired by Prime Minister Abdullah Ahmad Badawi and will provide tax breaks and matching grants to biotechnology companies. Its stated goal is to promote projects that can gain "international recognition".

This toned-down and yet outward-looking approach seems to be part of a more realistic

framework of education and science policies now being introduced. In Penang, for instance, the local government is establishing a research base that would include contract research activities in animal toxicology — which may be of interest to foreign companies. "Tests are cheaper here and the animal-rights issues are not as prominent," says Penang's mayor, Koh Tsu Khoo, who recognizes that investment in people will be essential. "We are building on brains rather than buildings," he says.

Rewarding merit

The central government is also taking steps to introduce more fundamental reforms. In 2003, English became the language of school instruction in maths and the sciences. Private universities have also been allowed — and are now providing opportunities for ethnic Chinese and Indian students who feel discriminated against by the state system. These include the Malaysian branch of Monash University, based near Melbourne in Australia. And officially, the rigid quotas used to enforce the 'Malays first' policy in higher education have given way to a merit-based system for allocating state university places.

But without standardized state university entrance exams, some critics remain sceptical about the likelihood of real progress. Unless Malaysia is able to shed its legacy of ethnic favouritism, they are dubious about the nation's chances of competing with its neighbours in biotechnology. "Frankly, while the government funds mostly Malays, it won't happen," says one foreign scientist based in Malaysia. "The government is putting a lot of money into biotech but I doubt that anything will come of it. I see a lot of white elephants."

David Cyranoski is *Nature's* Asian-Pacific correspondent.

Next week's *NatureJobs* will include a feature on Malaysia's private universities.



Marvels in the mist: Malaysia's biodiversity is a big draw for researchers — practically the only one.

What's the plot?

Can mathematicians learn from the narrative approaches of the writers who popularize and dramatize their work?

Sarah Tomlin is on the story.

"Mathematicians are quite shy, with very few exceptions," announced Pierre Cartier of the Institut des Hautes Études Scientifiques, south of Paris, opening his talk on mathematicians who have written autobiographies. "And most of them are in this room!" someone shouted from the floor. It was a rousing start to a highly unusual meeting.

Last month, a select group of about 30 mathematicians, playwrights, historians, philosophers, novelists and artists descended on the Greek island of Mykonos. The gathering aimed to find common ground between storytelling and mathematics, and was inspired by a profusion of books, movies and plays that, over the past decade, have dragged the subject out of a cultural wilderness.

Clearly, stories about mathematics have strong popular appeal. But what can professional mathematicians learn from the writers who are now taking an interest in their work? Can narrative approaches help the increasingly esoteric sub-fields of mathematics communicate with one another? And can such approaches help mathematicians frame the abstract problems that fill their working lives?

It was the hope of answering such questions that led the Greek novelist Apostolos Doxiadis to cook up the idea of bringing writers and mathematicians together. After graduating with a mathematics degree from New York's Columbia University at the age of 18, Doxiadis turned to his first loves of poetry and theatre. But in his late thirties, Doxiadis revisited mathematics with his 2001 novel *Uncle Petros and the Goldbach Conjecture*, in which the protagonist's attempt to prove this conjecture — that every even number is the sum of two primes — becomes a universal story about courage and suffering.

Excited by the narrative possibilities offered by mathematics, Doxiadis formed a foundation to further explore the theme — called Thales & Friends, after the first mathematician and philosopher in ancient Greece. And with financial backing from the Mathematical Sciences Research Institute in Berkeley, Califor-

nia, he set about organizing the Mykonos meeting. At first, the plan was for a smaller informal gathering, but it soon snowballed. "I'm surprised, frankly, at the response," Doxiadis told *Nature*.

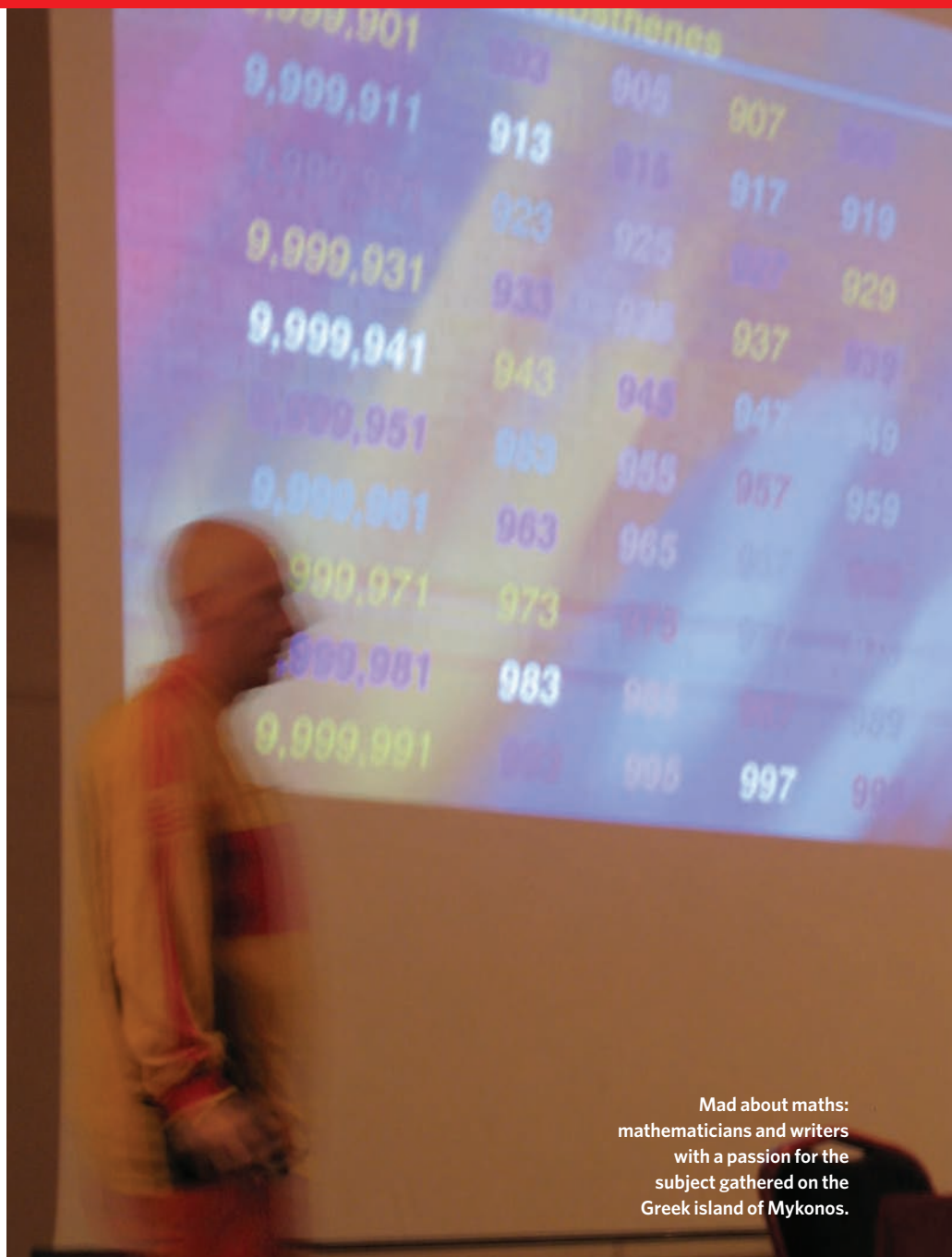
The venue was apt, given that ancient Greece was where the gulf between mathematics and story-telling first opened up. "Plato approved of mathematics, but despised poetry," says Rebecca Goldstein, a philosopher and novelist based in Hartford, Connecticut, who has used mathematicians as characters in several novels. Other participants blamed Euclid for introducing the impersonal, logical style that has characterized much mathematical writing ever since.

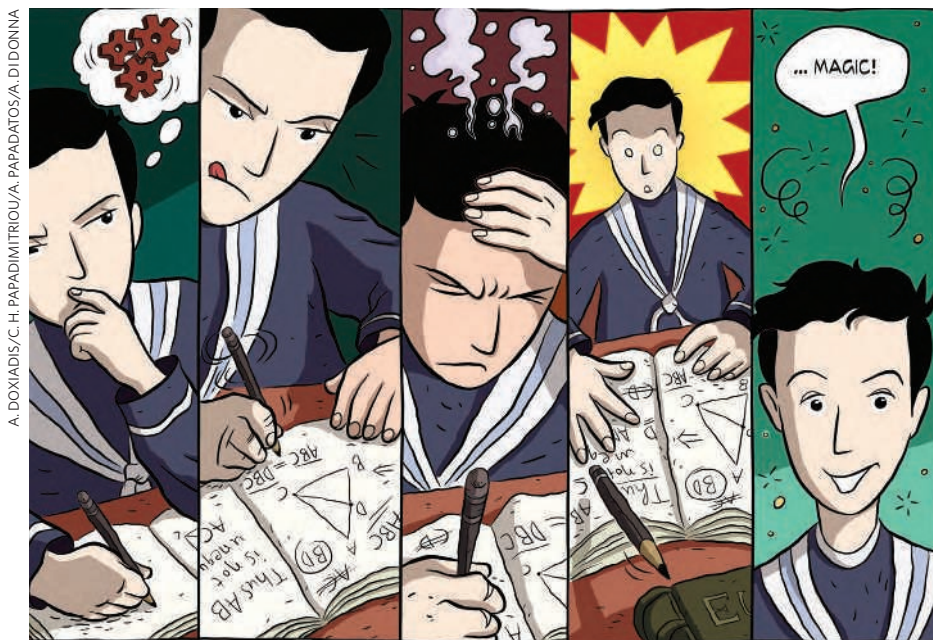
Whatever the historical reasons for the divide, it has been bridged with a vengeance in recent years. The spark came in 1995, with the solution of Fermat's last theorem by the British

mathematician Andrew Wiles. His 100-page proof made prime-time news, and was followed by a best-selling book, *Fermat's Enigma* by Simon Singh. Suddenly mathematics was fashionable. Next came a host of popular books on subjects such as zero, pi and irrational numbers, the Hollywood movie *A Beautiful Mind*, and plays such as David Auburn's *Proof* and John Barrow's *Infinities*. Today, there is even a US television series called *Numb3rs*, in which a detective relies on the skills of his mathematical-genius brother to solve crimes.

But if it is possible to tell stories about mathematics to a general audience, why do specialists in different branches of the discipline have so much difficulty communicating with one another? "Most mathematics papers are incomprehensible to most mathematicians," complains Tim Gowers of the University of Cambridge, UK, winner of a 1998 Fields

Mad about maths: mathematicians and writers with a passion for the subject gathered on the Greek island of Mykonos.





Brought to life: Alecos Papadatos' graphic novel has been inspired by the lives of logicians.

Medal — the nearest thing the subject has to a Nobel prize. "Publication has become just a formal stamp of approval — it is not a means of communication anymore," he adds.

Gowers is currently editing the *Princeton Companion to Mathematics*, which is scheduled to appear in 2006 and is intended to provide budding mathematicians with an accessible overview of the field. "Say you decided you wanted to do research in mathematics but you didn't know what area would appeal to you," says Gowers. "There is nothing available right now."

Tales of tables

Narrative approaches can help make arcane branches of mathematics more accessible to specialists, as well as to the lay public, Gowers argues. Persi Diaconis, a statistician at Stanford University in California, agrees. "I can only work on problems if there is a story that is real for me," he says. As one of the few applied mathematicians at the meeting, Diaconis perhaps has an easier time than most in applying narratives to his work. In *Mykonos*, he picked three of his papers, and told a story about each.

The first was a 1987 paper entitled 'Projection pursuit for discrete data', which deals with a mathematical technique for finding patterns in data in a systematic way. To make this technique come alive, Diaconis chose a real-life problem — the dating of Plato's most important texts. Scholars had previously classified the last five syllables of Plato's sentences as either 'short' or 'long', in the expectation that his writing style had changed over time, and that this would show up in these data. But they struggled to find anything meaningful. By looking at pairs of syllables using his technique, Diaconis was able to find patterns in the texts that had previously been hidden, and could subsequently work out an order for Plato's works.

Diaconis learned to love the abstraction of pure mathematics from his tutor Barry Mazur of Harvard University. But Mazur admits that he used to be mystified when Diaconis would ask: "I'm lost. What's the story?" Today, Mazur says he has woken up to the power of narrative, and in *Mykonos* gave an example of a 20-year unsolved puzzle in number theory which he described as a cliff-hanger. "I don't think I personally understood the problem until I expressed it in narrative terms," Mazur told the meeting. He argues that similar narrative devices may be especially helpful to young mathematicians, who seem particularly poor at explaining their work to others.

Mazur explained that he used narrative to help him develop a general organizing structure around the problem, which involves a conflict between theory and the large amount of data gathered on elliptic curves. Mazur did not find a solution by using the narrative device of a cliff-hanger, but it helped him to frame the question — and that, he argues, may be as important.

Mazur's explorations of narrative may sound trivial. And some of the writers who attended the *Mykonos* meeting admitted that they were left cold by this particular story. But Doxiadis explains that mathematicians, unlike experimental scientists, aren't used to dealing with conflicts between data and theory in this way. They like to understand a problem at its most basic level, from the inside. So for a mathematician, Mazur's depic-

"The *Mykonos* meeting marked the beginning of a rapprochement between the estranged arts of mathematics and story-telling."

tion of his number-theory problem in dramatic terms might be a genuine eye-opener. "The thing that inspired me was the idea that this could affect mathematics itself," says philosopher David Corfield of the Max Planck Institute for Biological Cybernetics in Tübingen, Germany.

Maths fever

In general, the non-mathematicians in *Mykonos* were passionate about the mathematical ideas that they have encountered. Alecos Papadatos is a cartoonist who is working on a graphic novel with Doxiadis and others about the history of logicians, most of whom died tragically. He claims he never liked mathematics, but now sees that the discipline is full of exciting stories. Barbara Oliver, who is artistic director of a theatre in Berkeley, California, echoes his view. She enjoyed directing the play *Partition* by Ira Hauptman — which is loosely based on the lives of two very different mathematicians, G. H. Hardy and Srinivasa Ramanujan — despite her lack of specialist knowledge. "I'm untutored, unskilled in mathematics," she says.

Whether mathematicians are similarly eager to embrace the joys of narrative remains unclear — particularly if telling a good story involves compromising standards of mathematical accuracy. Some of the most heated discussions in *Mykonos* were about the mistakes in a recent popular book on infinity, and whether they really matter. Diaconis, despite being an enthusiastic story-teller, still sees a sharp tension between narrative and mathematics. "To communicate we have to lie. If we don't we're deadly boring," he says. Diaconis waited until his 60th birthday to start writing a book on mathematics and magic, and he understands why others are cautious. "There's no reward for expository stuff, and a bias against it," he says.

Still, the *Mykonos* meeting at least marked the beginning of a rapprochement between the estranged arts of mathematics and story-telling. Another meeting is planned for next year — and there is much work to be done, if the tales of the thorniest mathematical problems are ever to be told. Take the Hodge conjecture: according to the website of the Clay Mathematics Institute in Cambridge, Massachusetts, which is offering a US\$1 million prize to anyone who can prove it, this asserts that "for particularly nice types of spaces called projective algebraic varieties, the pieces called Hodge cycles are actually (rational linear) combinations called algebraic cycles".

If that makes no sense to you, don't worry — most mathematicians regard the Hodge conjecture as 'unexplainable'. Let's hope, at least for the sake of the journalists who must try to communicate the essence of the conjecture to the general public if it is ever proved, that mathematicians will by then have honed their narrative skills.

Sarah Tomlin is a senior news feature editor for *Nature*.

BUSINESS

Fatal attraction

Oxford Instruments has paid dear for its bold efforts to stretch the boundaries of magnet performance, as **Andrea Chipman** reports.

For a British start-up in the physical sciences to establish clear global leadership in its market and a solid presence on the stock exchange is a rare thing indeed. It has taken more than 40 years for Oxford Instruments — the University of Oxford's very first start-up company — to get where it is today, with 1,200 employees and annual sales of £156 million (US\$276 million). But having made a name for itself as perhaps the world's leading supplier of high-specification magnets for research equipment, the company has suffered a few setbacks.

The firm was established in 1959 by a physicist, Martin Wood, who had just built Europe's first superconducting magnet. It grew steadily and was listed on the London Stock Exchange in 1983. But analysts say that its recent woes illustrate the pitfalls for companies whose scientific ambitions get in the way of profitability.

"They were always regarded as an extremely good business, but the results began to decline in the late 1990s," says one London-based analyst who has followed the company closely but declined to be identified. "Oxford Instruments was battling against the limits of its capabilities. Customers would decide on a number of specification changes as the system was being put together."

The problems originated largely in the company's superconductivity division, which produces superconducting magnets and low-temperature bench-top cryostats. It also manufactures superconducting wire, magnets for magnetic resonance imaging (MRI) and systems for a technique called ion cyclotron resonance, which determines the chemical composition of molecules.

The division's clients range from the health-care industry, which uses MRI for medical scans, to researchers at major laboratories, who use it to study materials and molecules.

Particle physicists, meanwhile, have been among the most demanding customers for Oxford Instruments' high-specification magnets, says Michael Cuthbert, sales manager for physical sciences at the company's US offices in Concord, Massachusetts.

The division's difficulties began after it took on a number of particularly demanding contracts on a fixed-cost basis, the London-based analyst says. "Because they were very challenging contracts, it was very appealing to take them

on for the kudos of doing them," she explains.

The main projects involved building magnets for a physics experiment at CERN, the European particle physics laboratory in Geneva, and for an MRI machine at the Pacific Northwest National Laboratory in Richland, Washington state.

CERN was planning an experiment called Compass, which sought to study the structure of hadrons using an existing synchrotron. The project incorporated several smaller studies, including one looking at the structure of protons and neutrons by scattering high-energy muons from them. This required the spins of all the particles to be polarized, using the field produced by a high-performance magnet.

Quenched field

Compass ordered the magnet in 1996, but physicists were never able to get it working at the required field strength, according to a source close to CERN. "The problem was that the magnet kept on quenching — the magnetic field could not be held for any length of time," said another source familiar with the project, referring to the release of liquid helium and subsequent loss of the field.

Ultimately, the CERN source said, the company gave up and made a substantial indemnity payment to the project. A number of former Oxford Instruments employees, working independently, eventually managed to get the equipment to work, the source added.

Around the same time, Oxford Instruments faced significant delays in delivering a high-specification MRI machine to the Pacific Northwest's Environmental Molecular Sciences Laboratory. The machine, which was



envisaged as the most powerful of its type in the world, was intended to help scientists study the structures of larger molecules and watch the interaction of molecules and cells at high resolutions (see *Nature* 383, 375; 1996).

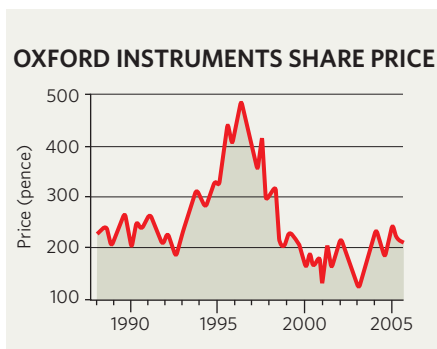
But it wasn't delivered until six years later, in 2002, and even then the laboratory held back a \$1.2-million final payment to be made once it was fully operational. Several current and former employees of the laboratory contacted by *Nature* declined to speak about the contract. An Oxford Instruments spokeswoman also declined to comment on the CERN and Pacific Northwest contracts.

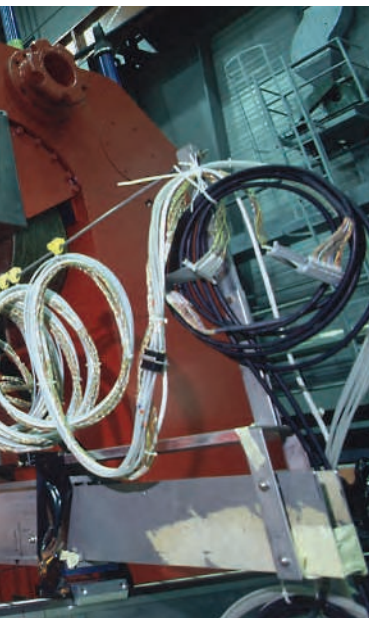
"In trying to be customer-focused and really challenging the technology, while perhaps not understanding all of the risks, we probably too often pushed to keep the customer satisfied," Cuthbert says.

As these contracts unfolded, the operating profit of the superconducting division fell from £10 million in 1997 to just over £1 million in 1999, followed by losses of £5.3 million in 2000.

In the wake of these problems, the division has tried to improve the way it manages exposure to risk, Cuthbert says. "We have regular management engineering risk reviews and assess all enquiries for risk well in advance of quotes," he said. "It's no good for the final customer or for any business to take on very high-risk projects and not succeed."

Over the past few years, the company has sought to rein in some of its technical ambitions and focus on the bottom line. But after staging something of a recovery, it reported some more problems last year. Quality issues had arisen at a plant making superconducting wire in the United States, and there had been a loss of business from a key client, Varian Medical





Damp squib: a magnet supplied by Oxford Instruments for CERN's Compass experiment struggled to produce a high enough field strength.

L. GUIRAUD/CERN

Systems of Palo Alto, California. In March, it appointed a new chief executive, Jonathan Flint — a trained physicist with management experience in other British high-tech firms.

A brand-new generation of magnets is a key part of the company's future growth strategy, says Adrian Philips, an analyst at the UK investment company Williams de Broe in Birmingham. The superconducting division has two particularly promising magnets in the pipeline, both of which look to extend the limits of its existing technology.

Molecular distinctions

A 950-megahertz, 22–23-tesla magnet will be delivered in October for use in MRI systems, Cuthbert says, allowing researchers to make more detailed distinctions between molecules. The division is also developing an 800-megahertz magnet that is actively shielded to restrain any stray field, improving safety and leading to more efficient use of lab space.

At the superconductivity division, analysts say, there will be less emphasis on ambitious, one-off projects to meet special customer requirements and more on standard products that play to the company's undoubted strengths.

Since Flint's appointment, the company's share price has remained relatively strong, and investors have high hopes for the future, says the London-based financial analyst who spoke anonymously to *Nature*.

"The company has always had undoubted technical expertise, but it has never been translated into consistent, profitable growth," she adds. "The arrival of a new chief executive from outside gives the hope that he will convert the technology into earnings growth, and deliver returns to shareholders." ■

IN BRIEF

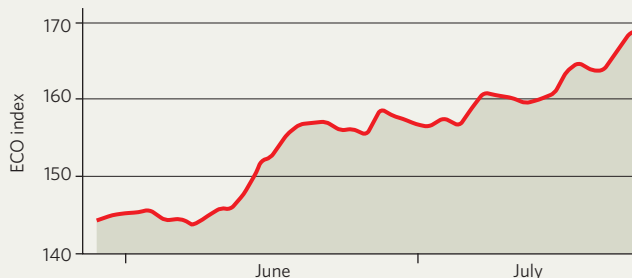
SUBSIDY PUSH The biotechnology industry and Elias Zerhouni, director of the National Institutes of Health, have called on the US government to allow companies funded by venture capital to rejoin a research-support programme. Lawmakers decided two years ago that companies majority-owned by venture capitalists should be kicked off the Small Business Innovation Research programme and that participation should be limited to small, independent businesses. But bills now before Congress would restore access to the subsidy for such companies.

INTEL PLANS FAB The world's largest chip manufacturer has announced that it will spend a cool \$3 billion on its next chip-fabrication facility near Phoenix, Arizona. Intel says this will be its first plant to produce chips with components only 45 nanometres apart — about twice as compact as current technology — on wafers 300 millimetres in diameter. As *Nature* went to press, Intel hadn't yet confirmed reports attributed to Israel's prime minister, Ariel Sharon, that it will also build a second chip-making plant in Israel.

AIDS AWARENESS A Chinese pharmaceutical company has signed a pact with a US health foundation to supply drug precursors at a discount price to companies that make generic AIDS drugs in India. The William J. Clinton Foundation, which was started by the ex-president in 2001, signed the agreement in Xiamen, China, with the Mchem Pharma Group. Under its terms, Mchem is also expected eventually to export finished AIDS drugs at affordable prices. It is the first Chinese company that the foundation has enlisted in its efforts to improve the availability of AIDS drugs in the developing world.

MARKET WATCH

CLEAN-ENERGY STOCKS



Stocks in companies dedicated to energy efficiency and renewables have surged forward this summer on mounting evidence that governments — and markets — are taking alternative energy more seriously.

The WilderHill Clean Energy Index (ECO on the American Stock Exchange) rose in value by almost 20% during June and July (see graph) as high oil prices, together with various corporate and political initiatives, bolstered investors' confidence in the sector.

The expected passage of an energy bill by the US Congress has fuelled positive sentiment, says Robert Wilder, a former political scientist at the University of California at Santa Barbara, whose company runs the index. Although the final bill won't include targets for renewable energy production, as some had hoped, its various measures have still boosted share values in the sector. "There are

enough subsidies in there for everyone," Wilder says.

He adds that the healthier performers in the index — such as Connecticut-based Distributed Energy Systems, which sells systems for supplying power from multiple sources into decentralized markets — are doing well on the basis of actual revenues and profits, rather than on speculation about future sales.

Michael Liebreich, head of London-based New Energy Finance and a member of the WilderHill index's advisory board, says an energy bill passed in China and the announcement of an environmental strategy by General Electric in May (see *Nature* 435, 410; 2005) have also improved market sentiment. The investment climate in clean energy, he says, "is moving from a gambling mentality to a more realistic phase, where risks are better understood by sophisticated investors".

SOURCE: WILDERSHARES

China building teams to tackle public-health crises

SIR — In his Commentary article “Is China prepared for microbial threats?” (*Nature* **435**, 421–424; 2005), David Ho called for China to train a cadre of public-health officers in programmes similar to those of the US Epidemic Intelligence Service. In fact, the Chinese Center for Disease Control and Prevention launched just such a programme in 2001: the Chinese Field Epidemiology Training Program (CFETP).

The CFETP is an international partnership with basic funding from the Chinese Ministry of Health and additional resources from the United Nations Children's Fund, the World Health Organization, the US Department of Health and Human Services and the US Centers for Disease Control and Prevention. It currently has 20 graduates, serving as field epidemiologists in 17 Chinese provinces, and offering guidance to 23 first- and second-year officers in 10 of these provinces.

On 19 February 2003, the CFETP sent six epidemiologists to Guangdong province to assist with surveillance and epidemiological investigation of the outbreak of atypical pneumonia now known as SARS. They demonstrated the effectiveness of personal protective measures for the prevention of transmission to hospital staff. When the outbreak spread to Beijing, the CFETP epidemiologists anchored surveillance efforts, investigated transmission chains and evaluated the effectiveness and utility of quarantine (J. Ou *et al. Morb. Mort. Week. Rep.* **52**, 1037–1040; 2003).

Since the SARS outbreak, the CFETP has undertaken more than 100 investigations on a wide range of public-health problems, including human influenza, HIV/AIDS, paratyphoid fever, measles, brucellosis, meningococcal meningitis, childhood injuries and disasters such as the 2004 typhoon in Zhejiang province. A CFETP officer and graduate recently travelled to Qinghai province to investigate the possibility of H5N1 avian influenza transmission to humans following the outbreak in migratory waterfowl; no such transmission was found.

Set up as a temporary programme, the CFETP is in the process of becoming a permanent unit of the Chinese Center for Disease Control and Prevention. In addition to the national effort, several provinces have developed their own programmes to extend training in field epidemiology and surveillance to their local health officers.

Building the CFETP to meet the needs of this vast and populous nation will require years of investment and training, as well as finding new ways to rapidly train and support health workers in the provinces. We look

upon Dr Ho's timely commentary as an opportunity to reinforce our activities.

Yu Wang*, Guang Zeng*, Robert E. Fontaine†

*Chinese Center for Disease Control and Prevention, 27 Nanwei Road, Xuanwu District, Beijing, 100050 China

†US Centers for Disease Control and Prevention, c/o US Embassy Beijing/CDC, PSC 461 Box 50, FPO AP 96521-0050, USA

Education and penalties are key to tackling misconduct

SIR — Your News story “One in three scientists confesses to having sinned” (*Nature* **435**, 718–719; 2005) identifies increasing pressure on scientists to publish papers and win grants as the main cause of misbehaviour. As a third-year graduate student, I think that the roles of the education and punishment systems may have been overlooked.

First, how many US graduate schools offer a compulsory ‘responsible conduct of research’ course? Young scientists, especially graduate students and postdocs, could easily misbehave because they lack education on the detrimental effects of misconduct.

Second, the low cost and risk associated with one misbehaviour may foster more misbehaviour. Graduate students and postdocs are usually the ones blamed when misconduct is revealed, while the professors tend to keep their positions and retain their funding.

If the benefits of misbehaving outweigh the possibility of being punished, academic misbehaviour is probably inevitable.

Kai Wang

Department of Microbiology, University of Washington, Seattle, Washington 98195, USA

Academia's 'misconduct' is acceptable to industry

SIR — As a physicist working in industry, I read the Commentary article “Scientists behaving badly”, by Brian C. Martinson and colleagues (*Nature* **435**, 737–738; 2005), with interest. The results are largely from scientists in academia. In my experience, some of the behaviours listed as “unacceptable” will seem quite normal to scientists working in industry.

Specifically, using someone else's ideas (misbehaviour no. 5) is regular commercial practice. Virtually all successful businesses are on the lookout for new ideas that can be applied within their own company for the purpose of gaining competitive advantage. If these ideas are not patent-protected, so much the better! Many companies employ scientists specifically to look for such opportunities.

Similarly, publishing the same data in multiple places (misbehaviour no. 11) is not

“Using someone else's ideas is regular commercial practice. If these ideas are not patent-protected, so much the better.”

— Ian Taylor

considered ethically dubious if there is no link between the number of publications and promotion prospects, which is generally the case in industry. It could be argued that publishing identical information at geographically dispersed conferences or journals is an aid to scientific communication. It is up to conference organizers and journal referees to police this behaviour if they do not like it.

Finally, withholding details of methodologies (misbehaviour no. 13) presents no ethical dilemma to scientists working in industry. When proprietary tests are developed to give competitive advantage, the results from such tests may be published, but the company should not be expected to divulge the underlying details.

Ian Taylor

8 Greenbank Road, Chester CH2 3RP, UK

Misconduct: pressure to achieve corrodes ideals

SIR — Your News story “One in three scientists confesses to having sinned” (*Nature* **435**, 718–719; 2005) and the corresponding Commentary article, “Scientists behaving badly”, by Brian C. Martinson and colleagues (*Nature* **435**, 737–738; 2005) fail to stress one important explanation for the differences in reported misbehaviours between mid-career and early-career researchers.

Most young scientists choose their career and engage in research with enthusiasm and idealism — that is, with the idea of ‘doing good’, which is essential for high-quality and ethical research. However, in the rough world of today's science, they are exposed to an environment in which impact factors and awards are more important than advancing the knowledge of mankind. They become prone to disillusionment and loss of vision.

The problem of disillusionment will not be solved by simply imposing sanctions on a broader range of misbehaviour. Science needs to regain a state in which researchers can maintain their idealistic motivation throughout their career. This goal can be reached only by the combined efforts of all parties participating in the scientific process.

Disclaimer: I am at an early stage in my scientific career and not yet disillusioned: the preceding statements are based on that perspective.

Lutz P. Breitling

Medical Research Unit, Albert Schweitzer Hospital, B.P. 118 Lambaréné, Gabon

BOOKS & ARTS

Cool is not enough

There's more to life than the second law of thermodynamics.

Into the Cool: Energy Flow, Thermodynamics and Life

by Eric D. Schneider & Dorion Sagan

University of Chicago Press: 2005. 362 pp. \$30, £21

J. Dooyne Farmer

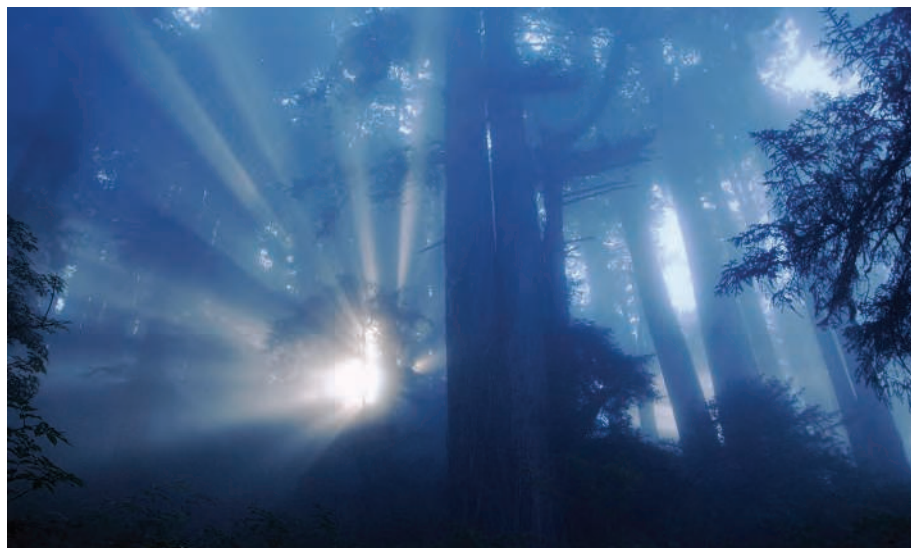
The level of organization in even the simplest living systems is so remarkable that many, if not most, non-scientists believe that we need to go outside science to explain it. This belief is subtly reinforced by the fact that many scientists still think the emergence of life was a fortuitous accident that required a good roll of the molecular dice, in a place where the conditions are just so, in a Universe where the laws of physics are just right.

The opposing view is that matter tends to organize itself according to general principles, making the eventual emergence of life inevitable. Such principles would not require any modifications of the laws of physics, but would come from a better understanding of how complex behaviour arises from the interaction of simple components.

Complex organization is not unique to living systems: it can be generated by very simple mathematical models, and is observed in many non-living physical systems, ranging from fluid flows to chemistry. Self-organization in non-living systems must have played a key role in setting the stage for the emergence of life. Many scientists have argued that certain principles of complex systems could explain the emergence of life and the universal properties of form and function in biology, and perhaps even provide insights for social science. The problem is that these principles have so far remained undiscovered.

In their book *Into the Cool*, Eric Schneider and Dorion Sagan claim that non-equilibrium thermodynamics provides the key principle that has been lacking. They review its application to topics ranging from fluid dynamics and meteorology to the origin of life, ecology, plant physiology, and evolutionary biology, and even speculate about its relevance to health, economics and metaphysics. The book contains a wealth of good references and is worth buying for this reason alone.

When the discussion sticks to applications where thermodynamics is the leading actor, such as the energy and entropy flows of the Earth, or the thermodynamics of ecological



G. JECAN/CORBIS

A complex problem: can a need to reduce energy gradients help to drive the evolution of forests?

systems, it is informative and worthwhile, but it is repetitive and seems disorganized in places.

The book is less successful as an exposition of a grand theory. It gets off to a bad start on the dust-jacket, which says: "If Charles Darwin shook the world by showing the common ancestry of all life, so *Into the Cool* has a similar power to disturb — and delight." While it may be wise to stand on the shoulders of giants, it is not advisable to stand back to back with one and call for a tape measure.

The authors' central thesis is that the broad principle needed to understand self-organization is already implicit in the second law of thermodynamics, and so has been right under our noses for a century and a half. Although the second law is a statement about increasing disorder, they argue that recent generalizations in non-equilibrium thermodynamics make it clear that it also plays a central role in creating order. The catchphrase they use to summarize this idea is "nature abhors a gradient". Being out of equilibrium automatically implies a gradient in the flow of energy from free energy to heat. For example, an organism takes in food, which provides the free energy needed to do work to perform its activities, maintain its form and reproduce. The conversion of free energy to entropy goes hand in hand with the maintenance of organization in living systems.

The twist is to claim that the need to reduce energy gradients drives a tendency towards

increasing complexity in both living and non-living systems. In their words: "Even before natural selection, the second law 'selects' from the kinetic, thermodynamic, and chemical options available, those systems best able to reduce gradients under given constraints." For example, they argue that the reason a climax forest replaces an earlier transition forest is that it is more efficient at fixing energy from the Sun, which also reduces the temperature gradient. They claim that the competition to reduce gradients introduces a force for selection, in which less effective mechanisms to reduce gradients are replaced by more effective ones. They argue that this is the fundamental reason why both living and non-living systems tend to display higher levels of organization over time.

This is an intriguing idea but I am not convinced that it makes sense. The selection process that the authors posit is never clearly defined, and they never explain why, or in what sense, it necessarily leads to increasing complexity. No one would dispute that the second law of thermodynamics is important for understanding the functioning of complex systems. Being out of equilibrium is a necessary condition for a physical phenomenon to display interesting complex behaviour, even if 'interesting' remains difficult to define. But the authors' claim that non-equilibrium thermodynamics explains just about everything falls flat. For

example, consider a computer. No one would dispute that a power supply is essential. Even for a perfectly efficient computer, thermodynamics tells us that it takes at least $kT \ln 2$ energy units to erase a bit, where T is the temperature and k is the Boltzmann constant. But the need for power tells us nothing about what makes a laptop different from a washing machine. To understand how a computer works, and what it can and cannot do, requires the theory of computation, which is a logical theory that is disconnected from thermodynamics. The power supply can be designed by the same person who designs them for washing machines.

The key point is that, although the second law is necessary for the emergence of complex order, it is far from sufficient. Life is inherently an out-of-equilibrium phenomenon, but then so is an explosion. Something other than non-equilibrium thermodynamics is needed to explain why these are fundamentally different. Life relies on the ability of matter to store information and to implement functional relation-

ships, which allow organisms to maintain their form and execute purposeful behaviours that enhance their survival. Such complex order depends on the rules by which matter interacts. It may well be that many of the details are not important, and that there are general principles that might allow us to determine when the result will be organization and when it will be chaos. But this cannot be understood in terms of thermodynamics alone.

Understanding the logical and physical principles that provide sufficient conditions for life is a fascinating and difficult problem that should keep scientists busy for at least a millennium. Thermodynamics clearly plays an essential part, and it is appropriate that the authors stress this — many accounts of the origin of life are easily rebutted on this point. But it isn't the principal actor; just one of many. The others remain unknown. ■

J. Doynne Farmer is at the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501, USA.

worked in a UK institute and at the World Health Organization (WHO). He repeatedly asserts that these illegals provided the Soviet Union with "significant" information.

Kouzmanov does provide some information on his agency's work. He describes how Westerners were targeted for recruitment by the KGB, and discusses the recruitment process and the means whereby data collected by agents and illegals were transported from the West to the Soviet Union. These procedures have previously been described by defectors and students of the Soviet intelligence system, and Kouzmanov's book adds little to the story already in the public domain. Disappointingly, it provides almost no information on how the KGB transformed the data into intelligence, and how this was then used.

According to Kouzmanov, individuals were deployed in the West and given numerous objectives related to spying on national programmes. For example, he describes a husband-and-wife team who, while operating a mock medical practice in Germany, were told by the KGB "to establish the locations of all NATO installations; their command personnel...air-force bases, and cruise-missile and rocket sites". It is doubtful that two individuals could accomplish all this. And Kouzmanov's explanation that the KGB placed agents in the WHO to obtain information about the "development of vaccines against the most dangerous human and animal viral diseases" seems rather lame, given that anyone could obtain this information simply by telephoning WHO representatives.

The author further alleges that around 1980 a KGB agent was placed inside the US Army Medical Research Institute of Infectious Diseases at Fort Detrick, Maryland, and that another agent was employed by an unnamed British institute (probably the National Institute for Biological Standards and Control, which was not engaged in biodefence). What did these agents do? Did they provide information about US and UK defensive efforts that might be used by the Soviet bioweapons programme? Did they inform their superiors that neither country actually had an offensive programme? Perhaps they provided information on the development of vaccines that might have been useful to the Soviet defensive programme?

In fact, Kouzmanov provides little information on the accomplishments of these and other agents in the biological field. Nor does he identify the Soviet research institutes with which the KGB allegedly collaborated in an effort to create more potent bioweapons, despite the fact that many of them are known today to Western security and academic communities.

Kouzmanov describes himself as a biophysicist with a microbiological background, so it is surprising how many technical mistakes he makes. For example, he misidentifies the bacteria *Bacillus anthracis* and *rickettsiae* as viruses, and misspells agents such as *Francisella tularensis* and *Yersinia pestis*.

Russia's secret weapons

Biological Espionage: Special Operations of the Soviet and Russian Foreign Intelligence Services in the West

by Alexander Kouzmanov

Greenhill: 2005. 192 pp. £12.99, \$19.95

Jens H. Kuhn, Milton Leitenberg & Raymond A. Zilinskas

In 1992, President Boris Yeltsin admitted that the former Soviet Union had supported a secret biological-warfare programme, in violation of the Biological Toxin and Weapons Convention, which the Soviet Union ratified in 1975. Some of the researchers and officials who operated the programme, such as Ken Alibek, Igor Domaradskii and Serguei Popov, have provided personal accounts that shed light on the clandestine system. However, the compartmentalization and secrecy so prevalent in the former Soviet Union mean that such accounts describe only a fraction of the nation's bioweapons programme. Almost nothing is known about the biological-warfare activities of the Soviet ministries of defence, health and agriculture, the security agencies and the national academies.

As a result, any new information on the roles of these agencies in the Soviet bioweapons programme is welcomed by those who are concerned about whether Russia is continuing with its bioweapons programme. This is the backdrop to the publication of a book by Alexander Kouzmanov, a former KGB agent, who claims to provide new and important information about the role of the KGB in the Soviet bioweapons programme. So, what do we learn from it?

Kouzmanov describes himself as a former employee of the top-secret Department 12 of



In the dark: the bioweapons programme run from KGB headquarters has remained largely secret.

Directorate S, the elite inner core of the KGB First Chief Directorate, which was responsible for operations abroad. One of the responsibilities of this department was to oversee 'illegals' — Russian intelligence operatives masquerading as Western nationals. Illegals were deployed to spy on Western biodefence activities, procure microbiological agents of interest for Soviet bioweapons research and development, and to perform acts of bioterrorism and sabotage. Kouzmanov was a case handler for several illegals, including some that allegedly

AP PHOTO/A. ZEMLIANCHENKO

Some of Kouzminov's claims are reiterations of stories that have been told before and have yet to be substantiated. For example, he alleges that the Soviets used *B. anthracis* and *F. tularensis* against German troops in the Second World War — an often-repeated story that has not been verified but has been discounted by microbiologists on the basis of epidemiological analyses. Kouzminov also asserts that Soviet agents obtained marburgvirus samples by exhuming victims of the first recorded outbreak of marburgvirus disease in Germany in 1967. But documentation of the official exchange of marburgvirus strains between German and Soviet microbiology institutes is publicly available.

Other claims, especially those in the final chapter, seem bizarre. If Kouzminov is to be believed, almost every outbreak of a new or

emerging infectious disease in the past 15 years — including the outbreak of foot-and-mouth disease in Britain in 2001 and the severe acute respiratory syndrome (SARS) pandemic in 2003 — may have been either a deliberate bioweapons attack or an accidental release of a genetically engineered microbe from a bioweapons facility. He also implies that the causative agents of hantavirus pulmonary syndrome were genetically engineered specifically to attack Native Americans. That allegations such as these would be made by a professional scientist in the face of a huge body of literature that seems to contradict them is astonishing.

It seems surprising that an insider can write a book about the special operations of Soviet foreign intelligence services in the West and provide so little about their achievements. At

best, *Biological Espionage* is the personal memoir of a former Soviet employee who writes about the practices of Soviet and Russian intelligence agencies in the biological field but provides little evidence of their accomplishments. Why was it written in the first place? If not to inform, then perhaps to misinform? ■

Jens H. Kuhn is a research scholar in the Department of Microbiology and Molecular Genetics, Harvard Medical School, Southborough, Massachusetts 01772, USA. Milton Leitenberg is at the Center for International and Security Studies, University of Maryland, College Park, Maryland 20742, USA. Raymond A. Zilinskas directs the Chemical and Biological Weapons Nonproliferation Program at the Center for Nonproliferation Studies, Monterey Institute of International Studies, Monterey, California 93940, USA.

Surface tensions

A reinterpretation, using damaged photographs, of a failed attempt to fly to the North Pole.

Colin Martin

Danish artist Joachim Koester recalls an ill-fated attempt to reach the North Pole by hydrogen balloon in his film installation at the 2005 Venice Biennale, *Message from Andrée*. It was inspired by the story of Swedish engineer Salomon August Andrée, whose balloon, optimistically named *The Eagle*, crash-landed in July 1897 a few hundred kilometres north of the Norwegian island of Spitsbergen, three days after taking off from Danes Island.

Andrée and his companions — another engineer, Knut Frænkel, and a physicist, Nils Strindberg — trekked across pack ice for three months, but failed to reach Franz Josef Land before the onset of winter. They died on White Island in mid-October. Their bodies and equipment were not found until August 1930, when an account of their misadventure, based on Andrée's diaries and Strindberg's notebooks, was published as *With the Eagle to the Pole*.

Strindberg had taken more than 100 pictures, using photographic plates with an expiry date of 1 January 1898. Remarkably, the exposed plates survived for 33 years and photographs were printed from them. Some, showing the aeronauts beside their damaged balloon and scenes from their trek, were re-touched to remove surface marks and were reproduced in the best-selling book.

Historians studying the photographs in the Andrée museum at Gränna also ignored this superficial layer of 'visual noise'. But Koester has used it to reinterpret the archive, by re-photographing many images with a 16-mm camera and producing a short



sequence of animated film. "If language defines our world, the black dots and light streaks on the photographs can be seen as bordering on the visible...pointing to the twilight world of what can be told and what cannot be told," he explains.

The jerky, random movements of the surface marks do not provide a definitive narrative, but Koester's film hints at the men's subconscious uncertainty, contrasting their innermost thoughts with the stoic 'public' thoughts conscientiously recorded in their journals.

In a catalogue essay on Koester's artistic response to the tragedy, Anders Kreuger characterizes nineteenth-century polar exploration as "a theatre of vanity and



GRÄNNA MUSEUM

Joachim Koester uses surface damage to photographic plates (left) to tell the tale of Salomon August Andrée's failed balloon flight to the North Pole.

delusion", inspired as much by Jules Verne's imagination as any defined scientific purpose. In retrospect, the notion of inexperienced aeronauts taking off into the Arctic wilderness in a hard to manoeuvre balloon, dependent on a strong southerly wind for their progress, seems foolhardy at best.

During their arduous trek across the pack ice, Strindberg coped psychologically by making endless lists: of the meals they ate, the equipment they carried, and, most poignantly, their ideas for improving future expeditions. In contrast, the incoherent short phrases and single words of Andrée's last diary entries have an abstract quality matched by Koester's film. The film, which can be seen in the Danish Pavilion at the Venice Biennale until 6 November, shifts the focus of Strindberg's images from the intentional to the accidental, reconfiguring a photographic narrative into a powerful affirmation of human endeavour in adversity. Colin Martin is a London-based writer.

NEWS & VIEWS

GEOCHEMISTRY

On the Moon as it was on Earth

Bernard Marty

Does the Moon's surface contain an archive of the early history of Earth? According to an intriguing idea, based on recently published analyses of lunar soils, it might do — and the proposal can be tested.

The onset of the terrestrial dynamo, resulting from slow cooling of the core and crystallization of its inner part, led to the development of the geomagnetic field that shields Earth's surface from extraterrestrial material and cosmic rays. The mechanism and timing of this event are unknown, but Ozima *et al.* (page 655 of this issue)¹ argue that the problem can be tackled from a fresh perspective.

Ozima *et al.* base their thinking on a new interpretation of geochemical data obtained from lunar soils recovered by the Apollo missions. They propose that, before the terrestrial dynamo kicked in, ions escaped from the top of Earth's atmosphere, and were implanted in lunar soils together with ions emitted by the Sun in the so-called solar wind. The process ended with the emergence of the geomagnetic field, and ions implanted in lunar soils have since been dominated by input from the solar wind. According to this model, then, the composition of volatile elements trapped in lunar soils is a mix of ancient terrestrial atmosphere and solar gas. The relative proportions varied through geological time, offering the prospect of investigating the epoch when the geodynamo started, as well as the compositions of these two sources.

Because it formed at high temperatures, the Moon as a whole is depleted in volatile elements such as those in the atmosphere and the oceans on Earth (hydrogen, carbon, nitrogen and the noble gases, for example). But these elements are abundant in lunar soils, pointing to the existence of extra-lunar sources, two of which are obvious.

The first stems from the fact that there is no plate tectonics on the Moon — after the initial stages of planetary building, the lunar surface evolved as a result of the continuous bombardment by 'planetary material'. This is material with a volatile-element composition different from that of the Sun, and not derived from the Sun, and includes asteroid and cometary bodies of various sizes. Some of the volatile elements trapped in these bodies were vaporized and re-trapped in the surface of soil grains. Second, ions from the solar wind are directly implanted into surface grains at characteristic depths of a few tens of nanometres.

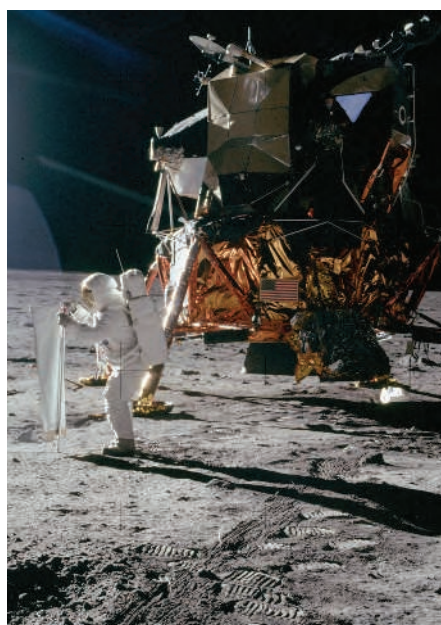


Figure 1 | Fieldwork on the Moon. Buzz Aldrin sets out the foil of a 'solar-wind collector'.

This component has been detected in both lunar soils and aluminium foils exposed to the Sun by the Apollo astronauts (Fig. 1). The isotopic compositions of helium, neon and argon (the only elements that could be analysed during the experiments) in both sets of samples are identical, showing that the solar-wind component dominates the noble-gas inventory of lunar soils.

But some other elements, in particular nitrogen, do not fit into this picture². Nitrogen has two isotopes, ^{14}N and ^{15}N , and their ratio in lunar soils varies by more than 30%, compared — in most situations — with less than 2% on Earth. No known nuclear process in the Sun or its corona can account for the nitrogen-isotope variations in lunar soils; moreover, nitrogen is enriched in lunar soils compared with noble gases in the solar wind, so it has been argued that the solar wind alone cannot account for the nitrogen-isotope and elemental abundances^{3,4}.

Five years ago, analyses of grains of lunar soil, made possible by technical advances, led to the proposal⁵ that the nitrogen component

of the solar wind is depleted in ^{15}N by more than 24% relative to terrestrial nitrogen. The suggestion was that the ^{15}N enrichment of lunar soil is 'planetary' in origin. Interplanetary dust particles are a particularly attractive candidate source in this respect — the flux of such particles dominates the extraterrestrial flux on Earth, and, compared with the solar wind, these particles are rich in ^{15}N .

Ozima *et al.*¹ have now come up with another interpretation. They propose that the ^{15}N -rich content of lunar soils came from the terrestrial atmosphere when the Moon was closer than at present and the geomagnetic field was not established, so allowing ions to escape from Earth. They base their hypothesis on isotope correlations resulting from mixing between two sources. On the one hand, a solar component free of deuterium (there is practically no deuterium in the Sun because it was consumed during the deuterium-burning stage of the star) and depleted in both ^{40}Ar (the solar gas is dominated by ^{36}Ar) and ^{15}N ; and, on the other hand, an atmospheric component containing deuterium (from the oceans) and ^{40}Ar (produced in the Earth by the decay of ^{40}K), and enriched in ^{15}N . Ozima *et al.* compute the flux of atmospheric species that their model demands and find it compatible with geochemical observations.

This interpretation can account for some of the noble-gas and stable-isotope variations, and it is consistent with the source strength required to account for the lunar soil inventory. But interplanetary dust particles are also enriched in deuterium and ^{15}N relative to the solar wind, and their flux, estimated from extrapolation of terrestrial measurements, can provide the required amounts of deuterium and nitrogen in lunar soils⁶ (the occurrence of ^{40}Ar is alternatively explained by degassing of the lunar interior). Another problem is that soils depleted in ^{15}N tend to show higher amounts of ^{40}Ar (relative to ^{36}Ar) than soils rich in ^{15}N , whereas Ozima and colleagues' proposal requires close association between ^{15}N -rich nitrogen and ^{40}Ar -rich argon. Lunar soils might in fact reflect the contributions of three sources, thus obscuring isotopic trends — a component from Earth's atmosphere, the

NASA JOHNSON

flux of which has varied drastically over time; a flux of interplanetary dust particles, which might have also varied through time; and a near-constant flux from the solar wind.

Ozima and colleagues' hypothesis is nonetheless thought-provoking, and has the merit of being testable. First, the light-element isotopic composition of the solar wind is still not well known. But progress in defining that will arise from an analysis of targets exposed for 27 months during the Genesis space mission. Second, as Ozima *et al.* point out, lunar soils from the far side of the Moon

should contain fewer atmospheric volatiles because that side has faced away from Earth for most of the Moon's history. We can hope that exploration of the far side will be a priority for future lunar missions; also, some meteorites may originate from the far side and are promising subjects for analysis.

We must wait and see. But if the hypothesis turns out to be correct, it will open new avenues of research into the history of Earth's magnetic field, and the possible connections with the evolution of life and with environmental conditions on the early Earth. ■

Bernard Marty is at the Centre de Recherches Pétrographiques et Géochimiques, 15 Rue Notre Dame des Pauvres, BP 20, 54501 Vandœuvre Cedex, France.

e-mail: bmarty@crpg.cnrs-nancy.fr

1. Ozima, M. *et al.* *Nature* **436**, 655–659 (2005).
2. Kerridge, J. F. *Rev. Geophys.* **31**, 423–437 (1993).
3. Geiss, J. & Boschler, P. *Geochim. Cosmochim. Acta* **46**, 529–548 (1982).
4. Wieler, R., Humbert, F. & Marty, B. *Earth Planet. Sci. Lett.* **167**, 47–60 (1999).
5. Hashizume, K., Chaussidon, M., Marty, B. & Robert, F. *Science* **290**, 1142–1145 (2000).
6. Marty, B., Hashizume, K., Chaussidon, M. & Wieler, R. *Space Sci. Rev.* **106**, 175–196 (2003).

NEUROBIOLOGY

Getting axons going

Juergen A. Knoblich

Neurons extend one long axon, through which they transmit electrical impulses to other cells in the nervous system. Surprisingly, it seems that where the axon forms is determined entirely within the neuron.

Neurons act as electrical relays: they collect information from other neurons through multiple extensions called dendrites, and transmit this information through one long protrusion, the axon. But the mechanism that determines where the axon forms at the neuronal surface has been unclear — it might be determined by an extracellular cue or by some intrinsic polarity that exists in the neuron even before the axon begins to grow. On page 704 of this issue, Calderon de Anda *et al.*¹ resolve this question by describing an unexpected correlation between axon outgrowth and the position of the centrosome — a structure that is involved in organizing the cell's internal scaffolding. The authors propose that where the axon forms is ultimately determined by the orientation of the neuron's

final cell division, after which it becomes fully specialized.

Neuronal differentiation can be followed in cell culture. When neurons from a brain area called the hippocampus are plated onto coated coverslips, they follow a stereotypical sequence of differentiation events². First, they form lamellipodia — highly dynamic protrusions that characteristically grow out of the leading edge of motile cells (stage 1). Shortly afterwards, they form four or five short extensions called neurites that have yet to gain the characteristics of axons or dendrites (stage 2). After 24 hours, one of the neurites extends rapidly and will become the axon (stage 3). The remaining neurites then acquire the characteristics of dendrites (stage 4), and finally, axons and dendrites form electrical contacts (stage 5).

What determines where the axon will form? To address this, Calderon de Anda *et al.*¹ analysed cultured rat hippocampal neurons from day 16 of embryonic development, immediately after the neurons' terminal division and before they become fully differentiated. Using real-time analysis, the authors show a sequence of events that correlates axon formation with a pre-existing polarity in the undifferentiated neuron (Fig. 1a). After the terminal division, the centrosome comes to lie opposite the plane of cleavage. During differentiation, the first lamellipodium forms in the region of cell membrane overlying the centrosome, and this is where the first neurite will grow. Calderon de Anda *et al.* also show that the axon consistently forms from the first neurite that grows out after the terminal division. So it is the plane of the terminal division that determines the site of axon emergence.

The results are not artefacts of cell culture because the authors find a similar correlation for hippocampal neurons differentiating *in situ* and for neurons in the developing eye of the fruitfly *Drosophila*. They are also not merely a correlation but a causal link, because ablation of centrosomes by a process known as chromophore-assisted light inactivation blocks axon outgrowth in cultured *Drosophila*

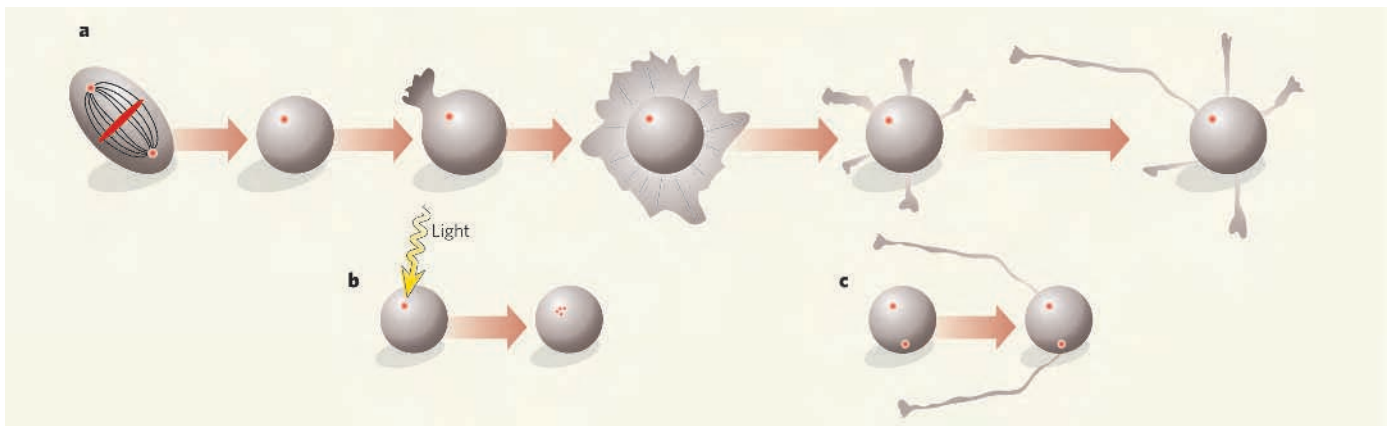


Figure 1 | Axon direction. Calderon de Anda *et al.*¹ show that the centrosome determines the position of axon outgrowth. **a**, Axon formation can be modelled using cultured hippocampal neurons. The centrosome comes to lie opposite the cleavage plane after the terminal cell division and initiates the formation of a lamellipodium, the first visible sign of neuronal differentiation. Many more lamellipodia then form, and finally the

first neurite extends from a position close to the centrosome. This first neurite ultimately forms the axon, and the other, later neurites become dendrites. **b**, No axon forms when the centrosome of a cultured *Drosophila* neuron is destroyed by laser light. **c**, When the final stages of cell division are inhibited, two centrosomes are present. They initiate the formation of two axons.

neurons (Fig. 1b). Furthermore, when the final stages of cell division are inhibited in hippocampal or cultured *Drosophila* neurons, the daughter cells contain two centrosomes. These cells form two long neurites that extend from positions directly overlying the two centrosomes (Fig. 1c). Although the experiments are carried out in different organisms, they indicate that centrosomes are required and sufficient for determining the position of axon outgrowth. Furthermore, they suggest the existence of a 'stage 0' in which cell polarity exists without any visible effect. This pre-existing polarity is used at later stages to direct neurite formation and axon specification.

These observations reveal exciting parallels between differentiating neurons and other cell types in which centrosomes initiate polarization. Shortly after fertilization, zygotes (single-celled embryos) of the nematode worm *Caenorhabditis elegans* become highly polarized along what will become the anterior-posterior axis³. This polarity is needed during the first cell division to segregate proteins differentially into what will become the two daughter cells, which will go on to have different fates. The axis of polarity in *C. elegans* zygotes is determined by the sperm entry-point, with polarization being initiated by an interaction between the centrosome (provided by the sperm) and the cell membrane³.

The similarity between *C. elegans* and neurons extends to the molecular level. Both polarity processes seem to involve an evolutionarily conserved set of proteins known as Par proteins^{3,4}. In *C. elegans*, Par-3 and Par-6 and the atypical protein kinase C (aPKC) localize to the anterior cell membrane, whereas Par-1 and Par-2 are concentrated posteriorly. In hippocampal neurons, Par-3 and Par-6 are found only in the axon, and if they are overexpressed, other neurites are induced to assume an axon-like morphology (see ref. 4 for a review). Moreover, vertebrate relatives of Par-1 seem to enhance neurite outgrowth. Although the distribution of Par proteins before differentiation now needs to be examined, the findings of Calderon de Anda *et al.*¹ suggest the existence of an evolutionarily conserved molecular machinery that polarizes cells and uses centrosome position as a reference point.

How does the centrosome influence the overlying cell membrane to induce neurite outgrowth at a particular position? The first morphological change during neuronal differentiation is the formation of lamellipodia. The process that creates these structures during cell migration is fairly well understood. It is driven by formation of a meshwork of actin proteins beneath the cell membrane. In fact, cell migration and neurite outgrowth might involve the same molecular machinery. It is remarkable that one of the first events in cell migration (at least in brain cells called astrocytes) is the reorientation of the centrosome to the future leading edge. This process is accompanied by a redistribution of the protein Cdc42 (a small GTPase,

involved in cell signalling) and — like axon formation and *C. elegans* polarity — it requires Par-6 and aPKC (ref. 5). Although the function and distribution of these proteins during early neuronal differentiation (the hypothetical stage 0) are unknown, it is likely that they control the cross-talk between centrosomes and the cell membrane in neurons as well.

What happens downstream of the Par proteins? The protein Rac (another small GTPase) is primarily responsible for lamellipodium formation, and Par-3 interacts with the Rac activators Tiam1 and STEF (ref. 4). So Par-3 could be responsible for lamellipodium formation through localized Rac activation. Thus, by analogy to other cell-polarity events, we can already draw a molecular pathway for neurite outgrowth that can be tested in the hippocampal neuron culture model. It should be noted, however, that *Drosophila* axon outgrowth is independent of Par-6 and aPKC, and the proposed pathway must be verified experimentally before any further conclusions can be drawn.

The results of Calderon de Anda *et al.*¹ imply

that the orientation of the final neuronal division is essential for correct wiring of the developing brain. Although such orientation is undoubtedly vital in invertebrates, its relevance in vertebrates is unclear⁶. The mechanisms responsible for the orientation of cell division have only recently begun to emerge in invertebrates³. Identification of those mechanisms in vertebrates will allow us to manipulate the orientation of cell division and test the effects on axon outgrowth — an experiment ultimately required to confirm the mechanism proposed by Calderon de Anda and colleagues. ■

Juergen A. Knoblich is at the Institute of Molecular Biotechnology, Dr Bohr Gasse 3-5, Vienna 1030, Austria.

e-mail: knoblich@imp.univie.ac.at

1. Calderon de Anda, F. *et al.* *Nature* **436**, 704–708 (2005).
2. Dotti, C. G. *et al.* *J. Neurosci.* **8**, 1454–1468 (1988).
3. Cowan, C. R. & Hyman, A. A. *Annu. Rev. Cell Dev. Biol.* **20**, 427–453 (2004).
4. Wiggan, G. R., Fawcett, J. P. & Pawson, T. *Dev. Cell* **8**, 803–816 (2005).
5. Etienne-Manneville, S. & Hall, A. *Cell* **106**, 489–498 (2001).
6. Wodarz, A. & Huttner, W. B. *Mech. Dev.* **120**, 1297–1309 (2003).

QUANTUM INFORMATION

Putting certainty in the bank

Patrick Hayden

A new way to manipulate quantum states resolves a long-standing conundrum about who knows what, and when and how, in the quantum world. The result is, as one has come to expect, startling and counterintuitive.

Claude Shannon's landmark 1948 theory of communication¹ tackles a nuts-and-bolts question: how do we find the best way to communicate using a given resource, such as a telegraph line or a satellite antenna? To answer that question, Shannon first took a detour into more philosophical territory by working out how to quantify the elusive concepts 'uncertainty' and 'information'. More than half a century on, quantum-information theorists have in many ways taken the opposite approach. Inspired by Shannon, but working with the notoriously counterintuitive theory of quantum mechanics, they seek to understand uncertainty and information in the quantum world by analysing the practical questions first, in the hope that the answers might then illuminate more fundamental conceptual issues.

On page 673 of this issue², Horodecki, Oppenheim and Winter demonstrate how effective this approach can be by justifying, in operational terms, a definition of conditional uncertainty that had previously been widely rejected owing to its strange and apparently nonsensical properties. In their formulation, what had been pathological becomes profound: with quantum information, it is

possible not just to be certain, but to be more than certain.

To understand what such a statement could mean requires first absorbing how information theorists think about uncertainty. Most readers will be able to decipher the following English sentence:

D r _ p / e _ e r _ / t h _ r d _ / e t _ e r .

As the omitted letters can be inferred with near-certainty from the others, they don't contribute to the uncertainty about the sentence and can therefore be compressed away. In general, the 'uncertainty' of a data source is the amount of space in bits required to transmit its output reliably.

Now consider another sentence, this time with more than three out of every four letters deleted:

T _ _ _ / _ _ _ / a _ _ _ r _ / _ _ / _ _ a _ .

It is no longer possible to decipher the sentence uniquely because there are many grammatically correct options. If some extra letters are provided, however, the task becomes feasible:

T _ _ s / i _ / a _ d _ r / t _ / r _ a d .



50 YEARS AGO

Recent investigations have shown that some of the free amino-acids in amphibian embryos change in concentration during early development. These changes are of interest because they may be related to the synthesis of new, specific proteins...it is possible to carry out analyses of the free amino acids not only in whole embryos but also in different regions representing different tissue primordia. Any regional differences in free amino-acid content which might be related to the early synthesis of tissue-specific proteins can in this way be detected. The results already quoted indicate that at least dorsoventral regional differences do exist.

From *Nature* 6 August 1955.

100 YEARS AGO

It is sometimes said that natural selection has ceased as regards civilised man; but very clearly this is an error. All civilised and most savage races are very stringently selected by various forms of zymotic disease. Thus in England practically everyone is brought into contact with the organisms which give rise to tuberculosis, measles and whooping-cough. Abroad, malaria, dysentery, and many other complaints play a similar rôle...The result of all this elimination by diseases demonstrates natural selection very beautifully. Every race is resistant to every disease strictly in proportion to its past experience of it...These facts appear to establish conclusively two truths, first that evolution is due solely to natural selection, and second that variations, except, perhaps, in rare instance, are not due to the direct action of the environment on the germ-plasm, but are "spontaneous." The Lamarckian doctrine is quite out of court. If ever acquirements are transmitted, it should be in the case of the profound and lasting changes affecting the whole body which result from disease; but in no instance is the effect produced by any disease on the race similar to that produced by it on the individual.

From *Nature* 3 August 1905.

The gap between what was provided at first (four letters), and what allowed us to decipher the sentence (no more than ten letters) illustrates the notion of 'conditional uncertainty' — the amount of extra information required to decipher a message.

One of Shannon's seminal results¹ was to find a simple formula for the uncertainty of a data source X . This function, usually written $H(X)$, is known as the Shannon entropy of X . Conditional uncertainty can be represented in similarly simple terms. If Y is used to represent the information already given to the receiver — the analogue of the indecipherable four letters in our example — the amount of extra information that must be provided is $H(X, Y) - H(Y)$, a quantity known as the conditional entropy of X given Y (ref. 3).

This second formula is easy to interpret: the extra information required is equal to the uncertainty in the total message, consisting of both X and Y , minus the uncertainty owing to Y alone, which should be subtracted, as Y is already known.

Among its many intuitive features, the conditional-entropy function is always greater than or equal to zero. That's because there is potentially more to be ignorant of in two messages X and Y together than in Y alone, so the inequality $H(X, Y) \geq H(Y)$ holds. For example, X and Y could represent future issues of the *Financial Times* and *The Wall Street Journal*, respectively: readers who take the time to follow both newspapers will be intimately familiar with the practical meaning of the inequality! In the context of conditional uncertainty, the interpretation is again highly intuitive: the amount of extra information required to decipher a message cannot be less than zero bits or, equivalently, it is impossible to be more than certain about the outcome of an event.

Intuitive indeed, but alas no longer true in quantum information theory. Horodecki, Oppenheim and Winter analyse² a quantum-mechanical version of the message-completion problem and find that the amount of extra information required can sometimes be less than zero qubits (a qubit is simply the quantum version of a bit). In their version of the problem, there are three participants: call them the sender, the receiver and the referee. The referee prepares a quantity of quantum information consisting of many particles, some of which he distributes to the sender and the receiver, and the rest he keeps for himself. The sender's job is to find an encoding that allows her to transfer her share of the information to the receiver using as few qubits as possible. To further isolate the quantum-mechanical features of the problem, the sender is also allowed to send old-fashioned messages consisting of bits at zero cost.

The authors show² that the number of qubits that the sender needs to transmit is precisely $S(A, B) - S(B)$, where A now refers to the sender's particles, B to the receiver's particles and S is the von Neumann entropy, a

direct quantum-mechanical generalization of Shannon's entropy. This formula is identical in form to the solution of the non-quantum version. With quantum particles, however, it is possible for A and B to be correlated in ways that are impossible in the classical situation⁴. In such cases, the systems A and B are said to be entangled (for a popular account of this phenomenon, see ref. 5). One consequence of entanglement is that conditional uncertainty, $S(A, B) - S(B)$, can sometimes be less than zero. In other words, the receiver can be more than certain!

In practice, if the receiver is more than certain, the sender doesn't need to transmit any qubits at all for the receiver to be able to decipher the message. So the receiver can put some certainty in the bank for a rainy day, in the form of extra entanglement with the sender that could be used to reduce the receiver's uncertainty about future messages. Entanglement is such a strong form of correlation that it can actually be used to send qubits from the sender to the receiver using a procedure known as quantum teleportation⁶. On the accounting ledger, therefore, having stored entanglement is almost as good as being able to communicate.

This neat and satisfyingly bizarre resolution disposes of a long-standing puzzle in quantum information theory: put simply, how to quantify who knows what. In more technical language, the puzzle was how to quantify conditional uncertainty. The formula $S(A, B) - S(B)$ had been proposed⁷, but was widely rejected because of its pathological tendency to become negative. Until now, no one had succeeded in finding a setting in which the formula's full range of positive and negative values would have a meaningful interpretation. (It was a quantum-information theorist's version of the famous conundrum from *The Hitchhiker's Guide to the Galaxy*: if 42 is the answer to Life, the Universe and Everything, what is the question?)

In addition to finally placing the quantification of uncertainty in quantum mechanics on a solid footing, the new result² opens the door to solving many previously intractable problems in quantum information theory. The authors provide a sampling of these applications in their paper, including an easy solution to a quantum version of the problem of many cellphones trying to communicate simultaneously to a single base station⁸. This astonishing solution shows that one sender's quantum information can, despite its fragility, be used to help decode the other senders' transmissions at higher rates than would otherwise be possible. Once again, quantum information has proved to be more versatile and more surprising than anyone expected. ■

Patrick Hayden is in the School of Computer Science, McGill University, 3480 University Street, Montreal, Quebec H3A 2A7, Canada. e-mail: patrick@cs.mcgill.ca

- Shannon, C. E. *Bell Syst. Tech. J.* **27**, 379–423, 623–656 (1948).
- Horodecki, M., Oppenheim, J. & Winter, A. *Nature* **436**, 673–676 (2005).
- Slepian, D. & Wolf, J. K. *IEEE Trans. Inform. Theory* **19**, 461–480 (1971).
- Bell, J. S. *Physics* **1**, 195–200 (1964); reprinted in Bell, J. S. *Speakable and Unsayable in Quantum Mechanics* (Cambridge Univ. Press, 1987).
- Aczel, A. D. *Entanglement: The Greatest Mystery in Physics* (Wiley, London, 2002).
- Bennett, C. H. et al. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
- Cerf, N. J. & Adam, C. *Phys. Rev. Lett.* **79**, 5194–5197 (1997).
- Yard, J., Devetak, I. & Hayden, P. preprint at <http://arxiv.org/abs/quant-ph/0501045> (2005).

ECOLOGY

Neutral theory tested by birds

Annette Ostling

A continental-scale analysis of habitat and bird distribution in South America provides the latest challenge for neutral theory — a controversial idea in ecology about what determines the make-up of communities.

How do different species end up living together in communities? Do they coexist only when each finds a different niche, or simply when they happen to disperse to the same habitable region? Debate over these questions intensified not long ago with the introduction of 'neutral theory'¹, a stochastic theory of community properties whose predictions have proven stubbornly robust, despite its disregard of the niches that many ecologists hold dear. Writing in *Proceedings of the National Academy of Sciences*, however, Graves and Rahbek² point out continental-scale patterns in the bird communities of South America that neutral theory may not be able to explain.

The dominant view in ecology is that species live together in communities only when they differ from one another. Species competing for the same nutrient or food source cannot coexist because one species will always be more efficient than the others and will quickly drive the rest to extinction³. Species that coexist must differ from one another in the resource they use most efficiently or in the environmental conditions to which they are best adapted — that is, they must have different niches. This view is often called 'niche-assembly'.

The contrary viewpoint is that communities are primarily shaped by historical accidents that influence where species disperse (a beetle floating to a distant island on flotsam, for example, or the uplift of a mountain range that blocks the flight of seeds from nearby forests). This view has deeper roots in evolutionary biology, where history is at centre stage, than in ecology, which concentrates on short-term interactions between species. The idea behind it is that, rather than being quickly out-competed, species that are less efficient at using a resource evolve to be as efficient as their competitors. The main criterion for coexistence is dispersal to the same habitable region. This view is sometimes called 'dispersal-assembly'.

The neutral theory tested by Graves and Rahbek² is the modern synthesis of dispersal-assembly into a mathematical framework. It

models a community as a finite collection of individuals that have identical probabilities of reproduction, death and dispersal. This yields predictions for community properties in terms of parameters that govern the stochastic changes that the community undergoes — for example the immigration or dispersal rate, the speciation rate, and the size of the community. Despite its simplicity, neutral theory's predictions have proven robust. Claims that it has been falsified⁴ have been followed by persuasive counter-arguments⁵.

Graves and Rahbek mount a new line of attack on neutral theory by testing it at the scale of an entire continent. Armed with an impressive bird-distribution database amassed from the collections of over 30 museums in 20 countries, as well as a global land-cover map created from satellite data, they quantify the correlation between the distribution of habitat

(or land-cover type) and the distribution of birds across South America at the resolution of 1° latitude by 1° longitude grid cells.

Put simply, Graves and Rahbek find that birds in more widespread habitats tend to be more wide-ranging. In particular, birds present in the lowland regions of the continent to the east of the Andes, where elevation changes slowly and habitats are widespread, are on average an order of magnitude more wide-ranging than those on the western side of the Andes, where topographic relief is at a maximum and habitat changes quickly in space (Fig. 1a). Furthermore, as one looks outwards from various locations on the continent, the change in bird-community composition is asymmetric and mimics the underlying changes in habitat (Fig. 1b).

Graves and Rahbek conclude that there is a strong causal influence of birds' habitat requirements on their spatial distribution across South America. They argue that this influence contradicts neutral theory, which ignores species differences in habitat requirements.

It is worth pointing out that the first of Graves and Rahbek's results, a correlation between habitat extent and species' spatial extents, could arise from a source other than habitat influence, a source that is instead consistent with neutral theory. The Andes act as a physical barrier to the dispersal of both birds and the flora that cover the landscape. This barrier, combined with the very different land areas available to dispersing species to the west and east of it, could alone explain the observed correlation.

But there are no dispersal barriers to explain the relationship between community composition and the distribution of riverside habitat evident in Fig. 1b. And further study of the

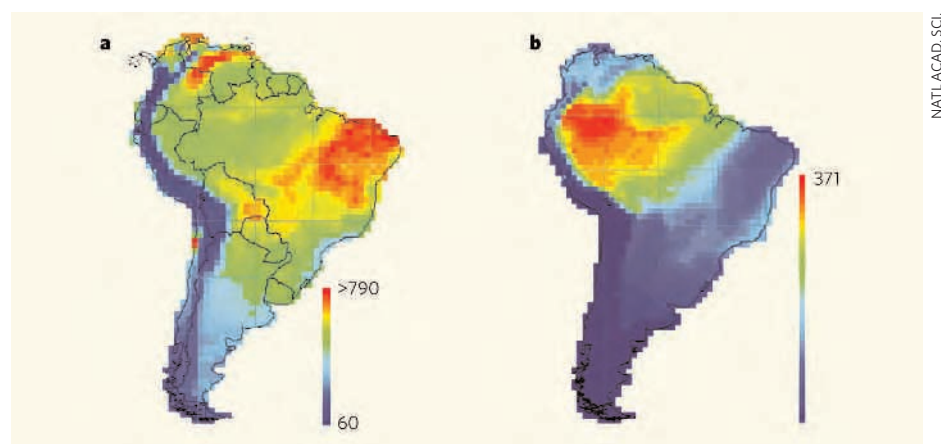


Figure 1 | Summary of Graves and Rahbek's results². **a**, Bird species in the lowland regions of South America, where habitat types are more widespread, are more wide-ranging than those on the western edge of the continent, where the Andes create quick changes in elevation and habitat type. Colours indicate median range-size in units of 1° latitude × 1° longitude cells. **b**, The composition of bird communities changes asymmetrically as one looks outwards from a location in the Amazon basin (here 1–2° S, 69–70° W), mimicking the underlying distribution of riverside habitat. Colours indicate the number of species in common with the focal location whose coordinates are listed. Neutral theory may be consistent with **a** but Graves and Rahbek are correct that it cannot predict the ecological importance of habitat evident in **b**. The theory may still be relevant at smaller scales, however, and species differences in habitat requirements can evolve under dispersal-assembly on a heterogeneous landscape. (Figures reproduced from ref. 2.)

NATL ACAD. SCI.

bird and land-cover databases would surely yield quantitative evidence of the influence of habitat. Practically all ecologists agree that species have habitat requirements that limit where they can live — tropical trees cannot survive on the Arctic tundra. Graves and Rahbek are correct that neutral theory cannot predict the resulting influence of habitat on community composition because it ignores species differences entirely.

But does the importance of habitat disagree with the letter or the intent of neutral theory? In other words, does it contradict the overall principle of dispersal-assembly?

Not necessarily. The idea of dispersal-assembly is not that differences between species do not exist — they are the inevitable result of disparate evolutionary histories. Rather, the idea is that species similarities, not their differences, lead them to find the same region habitable and to coexist. Neutral theory applies only in that realm of intermingling, where species are similar.

Habitat influence on species' distributions at any scale does indicate a role for niche-assembly, which has implications for ecological dynamics. The species that differ in the habitat they do best in cannot out-compete each other. Their differences allow them to coexist stably in the landscape.

However, unless habitat and species change in lock-step, habitat effects do not rule out a simultaneous role for dispersal-assembly. As Graves and Rahbek acknowledge, their observations limit only the spatial scale and groups of species within which neutral theory's unstable ecological dynamics may apply. Furthermore, differences between species in habitat requirements can arise from sources that are consistent with dispersal-assembly in a heterogeneous landscape over evolutionary timescales, such as from local selection for capabilities on a par with those of competitors. Selection for the avoidance of competition (or niche-assembly) may not be the evolutionary origin of these differences.

More empirical work is needed to distinguish between niche-assembly and dispersal-assembly on both ecological and evolutionary timescales. We also need to understand the implications of this distinction, and more refined ones, for judging the robustness and resilience of communities in the face of anthropogenic change. ■

Annette Ostling is in the Department of Ecology and Evolutionary Biology, Guyot Hall, Princeton University, Princeton, New Jersey 08544-1003, USA.
e-mail: aostling@princeton.edu

CANCER

Crime and punishment

Norman E. Sharpless and Ronald A. DePinho

Cellular senescence stops the growth of cells. This process, first glimpsed in cell culture, is now confirmed by *in vivo* evidence as a vital mechanism that constrains the malignant progression of many tumours.

Societies have traditionally taken three approaches to handling recidivist criminals: exile, execution and lifetime imprisonment. It seems that human cells use similar strategies to prevent rogue cells harbouring dangerous mutations from turning into fully fledged cancers. Epithelial tissue, such as that lining the airways and intestines, continuously renews and sloughs off, thereby sentencing some pre-cancerous cells to extra-corporeal exile. There is also a cellular version of the death penalty — apoptosis, a well-established anticancer mechanism. And in this issue, four groups^{1–4} report striking *in vivo* evidence that the body can subject potential cancer cells to the equivalent of a life-sentence: cellular senescence.

Senescence is a specific form of stable growth arrest provoked by diverse stresses, including the enforced expression of cancer-promoting genes in cultured cells. This 'oncogene-induced senescence' (OIS)⁵ is linked to known cancer pathways in cultured cells, notably the ARF–p53 and p16^{INK4a}–RB pathways (Fig. 1). But whether OIS is an authentic anticancer process *in vivo*, or simply an artefact of enforced oncogene expression in cells experiencing culture shock⁶, has been controversial.

This issue is settled by the new papers^{1–4}, which show that OIS occurs *in vivo* in several diverse precancerous tissues from both human and mouse. In addition, the work identifies much-needed markers of senescence, and further delineates the molecular underpinnings of this key tumour-suppressing process. A compelling feature of these studies is the consistency of OIS in response to a variety of cancer-causing mutations in different human tumour types and mouse-model systems. At the same time, the reports reveal that the molecular circuitry of OIS may be wired differently among tumour types.

Michaloglou *et al.* (page 720)¹ worked with cultures of human melanocytes (pigmented skin cells) and nevi (skin moles, the benign precursors of malignant melanoma). They found that nevi harbouring mutations of the BRAF protein (mutations that are frequently found in melanomas) have robust expression of senescence markers and do not seem to proliferate. In melanoma cells, however, senescence is extinguished and proliferation accelerated.

Curiously, the tumour suppressor p16^{INK4a} — a known activator of senescence that is deleted in melanoma cells — showed spotty

expression in nevi, and experimental depletion of p16^{INK4a} failed to increase BRAF-induced senescence in melanocyte cultures. Mutated BRAF in melanocytes also failed to induce the ARF and p53 tumour suppressors, two proteins integral to the activation of senescence in many systems. These results expose serious gaps in our understanding of the genes and pathways that function to constrain the transformation of nevi into lethal melanomas.

Exploring the evolution of prostate cancer, Chen *et al.* (page 725)² discovered senescence in early-stage prostate abnormalities in humans and in mice engineered to sustain prostate-specific deletion of the PTEN tumour-suppressor gene. However, in contrast to the situation in melanocytes, prostate OIS is dependent on p53, and co-deletion of PTEN and p53 cancelled senescence, promoting full-blown prostate cancer. Parallel studies using mouse models to dissect the role of the *Ras* oncogene in the lung and pancreas³ and in lymphoid cells⁴ reinforced similar principles. So, although previous work has established that the role of p53 as a tumour suppressor depends on its ability to mediate apoptosis, these papers emphasize that p53 can also mediate senescence in primary tumours.

Collado *et al.* (page 642)³ address a crucial need for better *in vivo* markers of OIS. So far, the gold standard has been the detection of an enzymatic activity associated with senescence (that of SA- β -gal)⁷. Although SA- β -gal has been used successfully to analyse human and mouse samples, this marker is not molecularly well-defined and demonstrates background activity in certain organs. Collado *et al.* employed an ingenious microarray screen to identify a small set of genes, the expression of which correlates strongly with senescence induced by the ERK protein. (ERK mediates the effects of certain cancer-causing mutations.) The correlations with gene expression are not seen when ERK is induced in the absence of senescence. These markers of OIS include protein-encoding genes and at least three RNA-encoding genes that are relevant to mouse tumour models of different tissues. These markers might predict OIS in precancerous abnormalities in humans.

Braig *et al.* (page 660)⁴ provide a penetrating biochemical view of senescence. Their experiments were guided by the observation of unusual foci of tightly packed DNA in senescent cells⁸. These foci possessed features of a form of silenced DNA called heterochromatin,

1. Hubbell, S. P. *The Unified Neutral Theory of Biodiversity and Biogeography* (Princeton Univ. Press, 2001).
2. Graves, G. R. & Rahbek, C. *Proc. Natl Acad. Sci. USA* **102**, 7871–7876 (2005).
3. Hardin, G. *Science* **131**, 1292–1297 (1960).
4. Clark, J. S. & MacLachlan, J. S. *Nature* **423**, 625–638 (2003).
5. Volkov, I., Banavar, J. R., Maritan, A. & Hubbell, S. P. *Nature* **427**, 696 (2004).

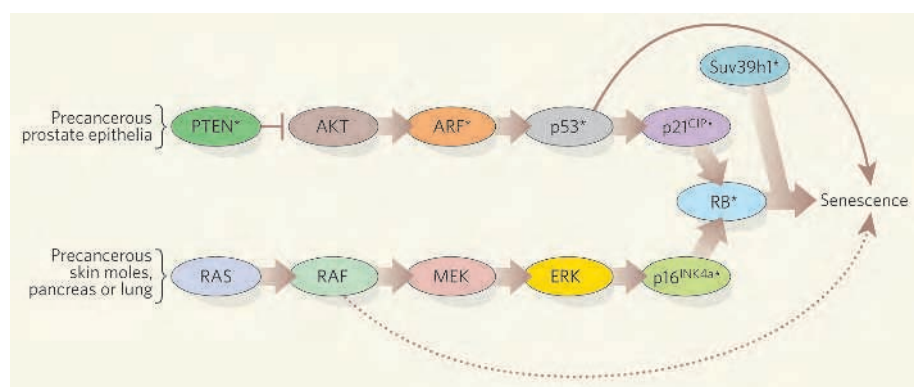


Figure 1 | A senescence wiring diagram. Oncogene-induced senescence has been linked to two major cell-signalling pathways that are often disrupted in cancer: the ARF–p53 and the p16^{INK4a}–RB pathways. Aberrant cancer-associated signals in premalignant cells activate these pathways and force would-be tumours into senescence, preventing progression to cancer. Although formerly thought to be molecularly homogeneous, senescence is now shown to differ depending on the cancer-promoting mutation and cell type. For example, ARF-independent senescence in human melanocytes (pigmented skin cells) may not require long-lasting expression of p16^{INK4a} or p53 (ref. 1), whereas ARF-induced senescence in the prostate occurs in response to mutation of AKT (ref. 2). Moreover, a biochemical dissection of the process shows that the Suv39h1 histone methyltransferase seems to act downstream of RB activation in RAS-induced senescence in lymphoma⁴. Asterisks denote proteins that are tumour suppressors.

and were characterized by specific methylation of some of the histone proteins around which the DNA is packed. The Suv39h1 protein is known to methylate histones, and binds physically to the RB tumour suppressor^{9,10}. So Suv39h1 seemed a reasonable candidate for a mediator of the senescence-promoting functions of RB.

Accordingly, Suv39h1 activity was shown to be required for *Ras*-induced OIS in lymphocyte cells. And in Suv39h1-deficient lymphomas, RB was not able to promote senescence as it normally does. These data suggest that Suv39h1 functions in concert with RB to alter DNA packaging in a manner that is required for senescence.

The findings of Braig *et al.* emphasize that OIS is an active process requiring specific molecular events that are sometimes perturbed, with malignant consequences. The authors reasoned that OIS would be inhibited by blocking histone deacetylases (HDACs) and DNA methyltransferases (DNMTs) because the activities of these enzymes seem to potentiate the ability of Suv39h1 to induce the formation of heterochromatin. Indeed, they found that lymphoma progression in a mouse model is accelerated by treatment with inhibitors of HDACs and DNMTs.

These findings are of added significance given the emerging application of 'epigenetic' therapies — targeted, for instance, at HDACs and DNMTs — in cancer and chronic diseases such as sickle-cell anaemia. The work raises the spectre that such therapies would provoke adverse consequences if they inhibited the senescence-mediated suppression of would-be cancer cells. Nonetheless, DNMT inhibitors have shown remarkable clinical activity in the bone-marrow disease myelodysplasia, and HDAC inhibitors have exhibited promising anticancer profiles in early trials.

Several issues regarding senescence remain to be addressed. Does senescence really mean life imprisonment for a precancerous cell, or is parole possible — that is, can senescence be reversed under some conditions? For example, could senescence in benign nevi be quelled, leading to melanoma, by further exposure to sunlight? The preliminary data here are mixed: some forms of senescence, such as that mediated by p53 in mouse cells^{11,12}, can be reversed by inactivation of p53 and/or RB, whereas other types, for example that mediated by p16^{INK4a} in human cells¹², cannot. Moreover, what eventually happens to senescent cells *in vivo*? Do they accumulate with age or are they culled through some unknown

mechanism? The fact that certain markers of senescence mount up with age suggests that senescent cells do accumulate to some extent.

Perhaps most importantly, these reports provide additional evidence that senescence is molecularly heterogeneous, requiring different pathways in different cell types, and in response to different oncogenic insults. A more precise molecular understanding of OIS is clearly needed. Just as apoptotic mechanisms are being used for therapeutic benefit, so a greater understanding of the stimuli that induce and enforce OIS will allow oncologists to exploit this crucial tumour-suppressor mechanism in cancer prevention and treatment.

Norman E. Sharpless is in the Departments of Medicine and Genetics, The Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599-7295, USA.

Ronald A. DePinho is in the Department of Medical Oncology, Dana-Farber Cancer Institute, and the Departments of Medicine and Genetics, Harvard Medical School, 44 Binney Street (M413), Boston, Massachusetts 02115, USA. e-mail: ron_depinho@dfci.harvard.edu

1. Michaloglou, C. *et al.* *Nature* **436**, 720–724 (2005).
2. Chen, Z. *et al.* *Nature* **436**, 725–730 (2005).
3. Collado, M. *et al.* *Nature* **436**, 642 (2005).
4. Braig, M. *et al.* *Nature* **436**, 660–665 (2005).
5. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. *Cell* **88**, 593–602 (1997).
6. Sherr, C. J. & DePinho, R. A. *Cell* **102**, 407–410 (2000).
7. Dimri, G. P. *et al.* *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
8. Narita, M. *et al.* *Cell* **113**, 703–716 (2003).
9. Nielsen, S. J. *et al.* *Nature* **412**, 561–565 (2001).
10. Vandel, L. *et al.* *Mol. Cell. Biol.* **21**, 6484–6494 (2001).
11. Sage, J., Miller, A. L., Perez-Mancera, P. A., Wysocki, J. M. & Jacks, T. *Nature* **424**, 223–228 (2003).
12. Beausejour, C. M. *et al.* *EMBO J.* **22**, 4212–4222 (2003).

EARTH SCIENCE

Trouble under Tonga?

George Helffrich

Earthquakes occur in cool, foundering tectonic plates deep within the Earth. But seismic data from the southwestern Pacific indicate that the minerals that make up the plates at depth don't behave as if they are cool.

Is there something wrong with our understanding of basic seismological features of Earth's mantle, the seismic discontinuities at depths of 410 km and 660 km? A report by Rigobert Tibi and Douglas A. Wiens, just published in the *Journal of Geophysical Research*¹, provides cause for thought about this possibility.

First, however, some context. Earth's surface consists of a mosaic of rigid tectonic plates, many of which have a hard life. They form at rock-melting temperatures at mid-ocean

ridges, sally across ocean basins and then sink out of sight in trenches, where they collide with another plate in a subduction zone. At the end of their life cycle, however, they leave a Cheshire-cat-like reminder of their former presence in the form of a sheet of earthquakes that extends, in many cases, from the surface down to about 700 km deep in the Earth. Here the grin of earthquakes fades.

The earthquakes in subducted plates happen because these plates are as much as 1,000 °C colder than their surroundings and

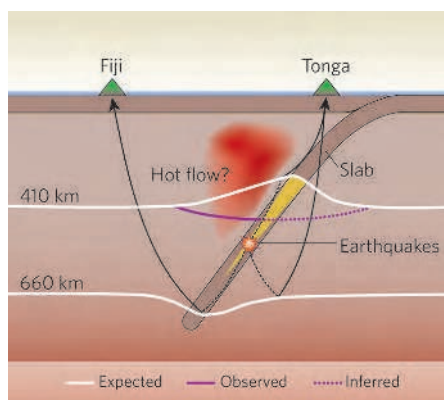


Figure 1 | Tibi and Wiens' observations¹ of the northern Tonga subduction zone. Earthquakes in the subducted slab emit two types of seismic wave, P (solid black lines) and S (dashed lines), which reflect off the 660-km discontinuity below the slab and travel back up to the surface to seismic stations on Fiji and Tonga. Upgoing S waves also interact with the 410-km discontinuity and convert to P waves. The time lags between the direct P waves from the quake (not shown for clarity) allow the depth of the discontinuities to be estimated. The '660' behaves as expected. But the '410' seems to be depressed (purple; dashed where not observed, but inferred), not raised (white), near the slab. A conjectural flow of hot mantle material might explain the depression here, and a possible region of metastable olivine inside the slab (yellow) might bend the 410 into a downwards wedge-shaped form, making it more difficult to observe. (Modified from Fig. 2 in ref. 1.)

are therefore brittle. They are, in essence, cold fingers stuck into the warm mantle and so provide a deep-Earth laboratory to probe phenomena that occur where the mantle is cooler than normal. One such phenomenon is the behaviour of the worldwide velocity jumps in seismic-wave speed at depths of 410 km and 660 km. For 30 years these discontinuities have been attributed to pressure-induced changes in mantle minerals², and, as such, have predictable responses to changes in temperature in subduction zones^{3–5}. Tibi and Wiens, however, conclude that they don't behave entirely as expected.

The problem is sketched in Fig. 1. A cold subducted plate should cause a mineralogical change related to the 410-km discontinuity (that of olivine to wadsleyite) to occur at lower pressure, and so shallower in the mantle. Conversely, the change related to the 660-km discontinuity, that of the higher-pressure form of wadsleyite, ringwoodite, to perovskite and magnesiowüstite, should occur at higher pressure, and so deeper. Careful analysis of subduction-zone earthquakes yields the depths of the '410' and '660' discontinuities from the timing of secondary signals arising from the reflection and conversion of seismic waves at the discontinuity boundaries.

Tibi and Wiens¹ deployed portable seismographic stations on Fiji and Tonga, two islands that bracket the Tonga subduction zone in the

southwestern Pacific, and looked for the signals from the 410 and 660 discontinuities in their data. This was a challenging project — observing conditions on ocean islands are extremely tough because of ocean-wave-generated seismic noise and problematic logistics. What Tibi and Wiens found, however, was that the 660 went down as expected but the 410 didn't go up everywhere, in part contradicting the mineralogical-change explanation for the discontinuities.

What is going on? Possibly two things. One, discussed by the investigators, is that hot mantle affected by the warm, ascending material that causes 'hotspot' volcanism in Samoa, nearby to the north, might be drawn southwards by the subduction beneath Tonga, so heating the vicinity of the slab. If it does, in my view there must be terrific thermal gradients in the slab, because parts of it are still cold enough to cause earthquakes. But this hot mantle might narrow the uplifted part of the 410 discontinuity, making it difficult to observe, and depress the 410 outside the slab. The observed 25-km depression corresponds to about 280 °C hotter temperatures⁵, not impossibly high for hotspots. Yet somehow the 660 behaves appropriately, suggesting that the thermal perturbation does not extend that deep or arose only recently, after the deeper parts of the slab had slipped past the present warm flow.

Alternatively, although Tibi and Wiens don't favour this explanation, the transition from olivine to wadsleyite could be hindered by the low temperatures in the slab, and be pushed downwards in the form of a metastable olivine wedge in the slab^{6,7}. If so, not all subduction zones behave in this way, because studies⁸ of such a zone near Japan found the

410 raised to 350 km by cold slab temperatures.

Before giving up on the explanation that the 410-km discontinuity is caused by mineralogical change, it would be worth checking regional Tongan topography with underside reflections^{8,9}. These are best investigated with observing stations that are at least 5,000 km distant, leading to reflection points that lie farther into the slab's interior because of the geometry. They are more sensitive to detecting topographic peaks than the methods used by Tibi and Wiens because reflection amplitudes remain strong for vertically travelling waves, whereas they decrease to zero for transmitted waves. This would settle the doubts about the dashed line in Fig. 1. On the other hand, the fading earthquake-limned grin of the hard-luck plate might be one of amusement at the head-scratching these observations are causing among Earth scientists. ■

George Helffrich is in the Department of Earth Sciences, University of Bristol, Wills Memorial Building, Queen's Road, Bristol BS8 1RJ, UK. e-mail: george@gly.bris.ac.uk

1. Tibi, R. & Wiens, D. A. *J. Geophys. Res.* **110**, doi:10.1029/2004JB003433 (2005).
2. Ringwood, A. E. *Composition and Petrology of the Earth's Mantle* (McGraw-Hill, New York, 1975).
3. Katsura, T. & Ito, E. *J. Geophys. Res.* **94**, 15663–15670 (1989).
4. Ito, E. & Takahashi, E. *J. Geophys. Res.* **94**, 10637–10646 (1989).
5. Bina, C. R. & Helffrich, G. R. *J. Geophys. Res.* **99**, 15853–15860 (1994).
6. Green, H. W., Young, T. E., Walker, D. & Scholz, C. H. *Nature* **348**, 720–722 (1990).
7. Kirby, S. H., Durham, W. B. & Stern, L. A. *Science* **252**, 216–225 (1991).
8. Collier, J. & Helffrich, G. *Geophys. Res. Lett.* **24**, 1535–1538 (1996).
9. Vidale, J. E. & Benz, H. M. *Nature* **356**, 678–683 (1992).

CELL BIOLOGY

Without a raft

Ben Nichols

The spatial organization of signalling proteins in the cell membrane is often ascribed to lipid-based 'rafts'. But single-molecule tracking reveals that such organization probably arises by protein-protein interactions.

Signal transduction — the relay of signals from outside a cell to inside — frequently involves bewildering patterns of interactions between several different types of protein at the cell surface. Work attempting to make sense of this complexity suggests that specific lateral organization of the interacting proteins in the membrane is key to their signalling functions. But how is such organization generated? Work by Douglass and Vale, published in *Cell*¹, begins to provide some answers, and emphasizes the utility of recently developed single-molecule imaging techniques in addressing the dynamic properties

of signalling networks. Moreover, the authors' experiments directly address the controversy over the mechanisms that generate localized variations in the composition of the cell membrane.

Simple mixtures of lipids in artificial membrane bilayers can segregate into regions that differ in the way the acyl chains of the lipids are packed together. This can spontaneously generate heterogeneity in the membrane, as lipids that prefer different local environments tend to separate out from one another. More-ordered acyl-chain packing is associated with the presence of increasing amounts of

cholesterol and sphingolipids — both found in natural membranes — and the resulting lipid domains tend not to be soluble in non-ionic detergents². Detergent-insoluble fractions enriched for particular proteins and lipids can also be isolated from cells. Extrapolating from the artificial membrane data, it has been proposed that these detergent-resistant fractions might be derived from functional domains, or 'lipid rafts', in cell membranes, where self-organization of the membrane lipids leads to the recruitment of specific proteins³. This lipid raft hypothesis has received much attention and is certainly appealing, but the correlation between detergent resistance and domain formation *in vivo* is a topic of some debate⁴. Artificial membranes may not be good models for cell membranes that are rich in protein and have two asymmetric layers of hundreds of different types of lipid.

The case of signalling through T-cell receptors is particularly germane to this debate. At the start of an immune response, T cells are activated when antigen molecules bind to the receptors on their surface. Stimulation of T cells usually occurs when stable contacts — referred to as 'synapses' — form between the T cell and so-called antigen-presenting cells⁵. There is a striking degree of spatial organization within the synapse; for example, molecules involved in the adhesion of the interacting cells, and activators and inhibitors of the signalling cascade, segregate and take on highly specific patterns^{5,6}. Moreover, because several of the proteins recruited to the T-cell synapse are found in detergent-resistant membrane fractions, various functions have been ascribed to lipid rafts in organizing these proteins during T-cell-receptor signalling⁷. This instance of a signalling machine in which the appropriate spatial distribution of its components is likely to be central to its function is a promising place to look for direct evidence of a physiological role for lipid rafts.

Douglass and Vale¹ adopted a widely used method for stimulating T cells — cross-linking the receptors using specific antibodies. But they modified it so as to be able to view the stimulated surface of the cell using either conventional confocal microscopy or total internal reflection imaging; the latter is a highly sensitive approach that can detect single fluorescent molecules in a narrow focal plane (Fig. 1). This remarkable technical achievement allowed individual fluorescent proteins involved in T-cell-receptor signalling to be tracked directly. Several of these proteins clustered together, for example the stimulatory co-receptor CD2, the adaptor protein LAT and the enzyme Lck; the negative regulator CD45 did not occur in the clusters.

LAT seems to play a prominent role in generating clustering, as the proteins do not group together in cell lines lacking LAT. By using a laser to photo-bleach a spot on the membrane, the researchers followed fluorescent molecules as they moved into the bleached area. This

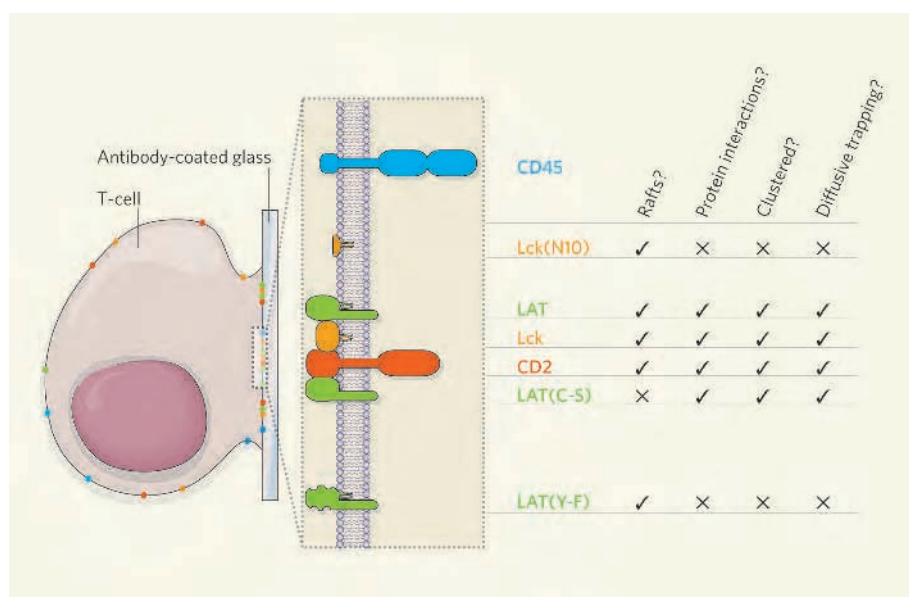


Figure 1 | Clustering to signal. Incorporation of proteins into lipid rafts is not related to the dynamic clustering that occurs during activation of the T-cell receptor, according to work by Douglass and Vale¹. The authors allowed T cells to settle on cover-slips coated with antibodies that cause localized activation of T-cell-receptor signalling. A combination of confocal microscopy and single-molecule imaging revealed that the proteins LAT, Lck and CD2 group together, and that these clusters exclude the negative regulator CD45. Freely diffusing LAT and Lck molecules can become transiently trapped in the cluster regions of the membrane. Clustering and this diffusive trapping require the protein-protein interaction domains of LAT and Lck (which are missing or defective in the LAT(Y-F) and Lck(N10) proteins). However, these actions are not affected by mutations that abrogate incorporation into lipid rafts (LAT(C-S)).

showed that there was a constant exchange of individual protein molecules in and out of clusters that were themselves stable over time. Finally, ingenious experiments tracking individual LAT, Lck and CD45 molecules relative to clusters containing CD2 revealed that these clusters can temporarily trap freely diffusing LAT and Lck, but not the negative regulator CD45.

Douglass and Vale's system provided an excellent opportunity to unravel the mechanisms that generate the spatial organization of signalling molecules (Fig. 1). A mutant of LAT, which lacks the acylated amino-acid residues required for incorporating the protein into lipid rafts⁸, showed identical clustering and diffusional trapping to normal LAT. Conversely, a LAT mutant lacking residues required for specific protein-protein interactions did not cluster or show diffusional trapping. Similar results were obtained with a truncated form of Lck that retains only the portions required for membrane association and raft incorporation. This mutant also did not cluster or show diffusional trapping.

These data do not support a link between raft incorporation (detergent resistance) and spatial organization during T-cell-receptor signalling. The authors propose instead that the clustering and diffusional trapping are best explained by a network of protein-protein interactions between the relevant signalling molecules. Although such a network remains to be characterized directly, previous studies corroborate the primary role of protein-

protein as opposed to protein-lipid interactions in T-cell-receptor signalling⁹. They also do not support the proposed link between raft incorporation and protein recruitment to membrane regions where the T-cell receptor has been activated¹⁰.

Now that our picture of protein dynamics during spatially organized signalling is beginning to reach single-molecule resolution, there seems no need to invoke the raft model to explain these dynamics. However, before summarily wielding Occam's razor, we should remember that model systems such as that used by Douglass and Vale do not fully replicate the properties of the T-cell synapse *in vivo*, and further subtleties doubtless remain to be investigated.

Ben Nichols is at the Laboratory of Molecular Biology, Division of Cell Biology, Hills Road, Cambridge CB2 2QH, UK.
e-mail: ben@mrc-lmb.cam.ac.uk

1. Douglass, A. D. & Vale, R. D. *Cell* **121**, 937-950 (2005).
2. London, E. & Brown, D. A. *Biochim. Biophys. Acta* **1508**, 182-195 (2000).
3. Simons, K. & Ikonen, E. *Nature* **387**, 569-572 (1997).
4. Munro, S. *Cell* **115**, 377-388 (2003).
5. Davis, D. M. & Dustin, M. L. *Trends Immunol.* **25**, 323-327 (2004).
6. Kupfer, A. & Kupfer, H. *Semin. Immunol.* **15**, 295-300 (2003).
7. Harder, T. *Curr. Opin. Immunol.* **16**, 353-359 (2004).
8. Zhang, W., Triple, R. P. & Samelson, L. E. *Immunity* **9**, 239-246 (1998).
9. Hartgroves, L. C. *et al. J. Biol. Chem.* **278**, 20389-20394 (2003).
10. Glebov, O. O. & Nichols, B. J. *Nature Cell Biol.* **6**, 238-243 (2004).

BRIEF COMMUNICATIONS

Dogs cloned from adult somatic cells

Two Afghan pups could help to unravel the genetics behind the assorted traits of other canine breeds.

Several mammals — including sheep, mice, cows, goats, pigs, rabbits, cats¹, a mule², a horse³ and a litter of three rats⁴ — have been cloned by transfer of a nucleus from a somatic cell into an egg cell (oocyte) that has had its nucleus removed. This technology has not so far been successful in dogs because of the difficulty of maturing canine oocytes *in vitro*. Here we describe the cloning of two Afghan hounds by nuclear transfer from adult skin cells into oocytes that had matured *in vivo*. Together with detailed sequence information generated by the canine-genome project^{5,6}, the ability to clone dogs by somatic-cell nuclear transfer should help to determine genetic and environmental contributions to the diverse biological and behavioural traits associated with the many different canine breeds^{7,8}.

Successful somatic-cell nuclear transfer (SCNT) depends on the quality, availability and maturation of the animal's unfertilized oocytes. Unlike other mammals, dogs ovulate at first meiotic prophase, and their oocytes mature for 2 to 3 days in the oviduct's distal regions. Previously, intra- and interspecific canine embryos have been constructed by canine SCNT into canine and bovine oocytes, respectively, but this did not result in viable offspring⁹.

We collected oocytes matured *in vivo* at metaphase II about 72 hours after ovulation by flushing the oviducts. (For details of methods, see supplementary information.) Donor fibroblasts were obtained from an ear-skin biopsy of a male Afghan hound and cultured for two to five passages (in which fully grown cells are transferred to a new culture dish). For SCNT, the chromosomes of the unfertilized canine oocytes were removed by micromanipulation, and a single donor cell was transferred into each enucleated oocyte. The couplets were fused and only successfully fused couplets (75%) were activated. The activated oocytes were then transferred into the oviducts or uterine horns of recipient dogs at times appropriate to the embryos' developmental stages. We collected an average of 12 oocytes from each female, and a total of 1,095 reconstructed canine embryos were transferred into 123 recipients.

Three pregnancies were confirmed by ultrasound scans at 22 days' gestation in recipients after transfer of constructs. Pregnancy was established only after embryo transfer of very-early-stage nuclear-transfer constructs (that is,



Figure 1 | Dog cloned by somatic-cell nuclear transfer. **a**, Snuppy, the first cloned dog, at 67 days after birth (right), with the three-year-old male Afghan hound (left) whose somatic skin cells were used to clone him. Snuppy is genetically identical to the donor Afghan hound. **b**, Snuppy (left) was implanted as an early embryo into a surrogate mother, the yellow Labrador retriever on the right, and raised by her.

less than 4 hours after oocyte activation). This transfer of early-stage embryos is a crucial factor in successful assisted reproductive technology for dogs. One fetus miscarried and two others were carried to term.

We named the first cloned dog Snuppy (for Seoul National University puppy); it is shown in Fig. 1a with the male Afghan fibroblast donor. Snuppy was delivered by caesarian section after 60 days (full term) from a yellow Labrador surrogate mother (Fig. 1b); his birth weight was 530 g. The second SCNT dog, NT-2, was carried by a mixed-breed surrogate, and was also delivered at 60 days, weighing 550 g (normal range for Afghans in a litter, 482–680 g). He experienced neonatal respiratory distress during the first week, seemed to recover, but died on day 22 as a result of aspiration pneumonia; no major anatomical anomalies were evident post mortem.

We tested whether the cloned dogs were genetically identical by microsatellite analysis of genomic DNA from the donor Afghan, the cloned dogs and the surrogates (see supplementary information). Analysis of eight canine-specific microsatellite loci confirmed that the cloned dogs were genetically identical to their donor dog. However, the efficiency of cloning is still very low (2 dogs from 123 recipients, or 1.6%) compared with the rates for cats¹ and horses³.

In addition to the benefits that cloning technology may generally provide (the preservation of rare species and therapeutic cloning

— once canine embryonic stem cells become available), this technology could become a useful research tool for studying the genetics of outcrossed populations.

Byeong Chun Lee*, **Min Kyu Kim***, **Goo Jang***, **Hyun Ju Oh***, **Fibrianto Yuda***, **Hye Jin Kim***, **M. Hossein Shamim***, **Jung Ju Kim***, **Sung Keun Kang***, **Gerald Schatten†**, **Woo Suk Hwang***

*Department of Theriogenology and Biotechnology, College of Veterinary Medicine, Seoul National University, Gwanak-gu, Seoul 151-742, South Korea
e-mail: hwangws@snu.ac.kr

†Pittsburgh Development Center, Magee-Womens Research Institute, Departments of Obstetrics–Gynecology–Reproductive Sciences and Cell Biology–Physiology, University of Pittsburgh School of Medicine, Pittsburgh, Pennsylvania 15213, USA

1. Shin, T. *et al. Nature* **415**, 859 (2002).
2. Woods, G. L. *et al. Science* **301**, 1063 (2003).
3. Galli, C. *et al. Nature* **424**, 635 (2003).
4. Zhou, Q. *et al. Science* **302**, 1179 (2003).
5. Ostrander, E. A. & Giniger, E. *Am. J. Hum. Genet.* **61**, 475–480 (1997).
6. Sutter, N. B. & Ostrander, E. A. *Nature Rev. Genet.* **5**, 900–910 (2004).
7. Lohi, H. *et al. Science* **307**, 81 (2005).
8. Modiano, J. F. *et al. Cancer Res.* **65**, 5654–5661 (2005).
9. Westhusin, M. E. *et al. J. Reprod. Fert. Suppl.* **57**, 287–293 (2001).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/436641a

TUMOUR BIOLOGY

Senescence in premalignant tumours

Oncogene-induced senescence is a cellular response that may be crucial for protection against cancer development^{1,2}, but its investigation has so far been restricted to cultured cells that have been manipulated to over-express an oncogene. Here we analyse tumours initiated by an endogenous oncogene, *ras*, and show that senescent cells exist in premalignant tumours but not in malignant ones. Senescence is therefore a defining feature of premalignant tumours that could prove valuable in the diagnosis and prognosis of cancer.

We used a mouse model for cancer initiation in humans: the animals have a conditional oncogenic *K-rasV12* allele that is activated only by the enzyme Cre recombinase³, causing them to develop multiple lung adenomas (pre-malignant tumours) and a few lung adenocarcinomas (malignant tumours). Senescence markers previously identified in cultured cells were used to detect oncogene-induced senescence in lung sections from control mice (expressing Cre) and from *K-rasV12*-expressing mice (expressing Cre and activated *K-rasV12*). We analysed p16^{INK4a}, an effector of *in vitro* oncogene-induced senescence¹, and *de novo* markers that we identified by using DNA microarray analysis of *in vitro* oncogene-

induced senescence (see supplementary information). These *de novo* markers are p15^{INK4b} (also known as CDKN2B), Dec1 (BHLHB2) and DcR2 (TNFRSF10D). In addition, we looked for two features evident in cultured senescent cells, namely the expression of senescence-associated β -galactosidase⁴ and the presence of senescence-associated heterochromatin foci⁵.

Staining with antibodies against p16^{INK4a}, p15^{INK4b}, Dec1 and DcR2 revealed abundant positive cells in adenomas, whereas adenocarcinomas were essentially negative (Fig. 1a). By contrast, the proliferation marker Ki-67 revealed a weak proliferative index in adenomas compared with adenocarcinomas (Fig. 1a). Lung cryosections from *K-rasV12* mice stained for senescence-associated β -galactosidase gave an intense signal in the adenomas, whereas adenocarcinomas gave a weak or negative signal (Fig. 1b). Adenomas were also strongly positive for HP1- γ , which indicates the formation of senescence-associated heterochromatin foci⁵, whereas adenocarcinomas were negative (Fig. 1c). These results were consistently found when using *K-rasV12* mice carrying Cre transgenic alleles that were expressed either inducibly by tamoxifen

(Cre-ER) or constitutively (CMV-Cre)³; 5–10 mice were analysed for each marker. The results have also been confirmed by immunoblotting and by quantitative real-time polymerase chain reaction with reverse transcription (see supplementary information; these analyses also included p19^{ARF}, an effector of senescence-associated β -galactosidase⁶).

Extending these observations, we combined the *K-rasV12* allele with a transgenic Cre allele that targets the expression of the oncogene to the pancreas. These compound mice develop premalignant lesions (pancreatic intraductal neoplasias) that progress into malignant ductal adenocarcinomas (C.G., A.J.S. and M. Barbacid, unpublished results). As in lung adenomas, these premalignant lesions were positive for our oncogene-induced senescence markers, whereas ductal adenocarcinomas were negative. Similarly, chemically induced skin papillomas, which harbour *H-ras* oncogenic mutations, were also positive for oncogene-induced senescence markers (see supplementary information).

We infer from our findings that oncogene-induced senescence may help to restrict tumour progression. In most cells, oncogenic *K-ras* signalling is attenuated and is therefore not sufficient to trigger tumour formation or senescence^{3,7}. Presumably, rare cells that do not fully attenuate oncogenic Ras are at the origin of both premalignant and malignant tumours. We conclude that a substantial number of cells in premalignant tumours undergo oncogene-induced senescence, but that cells in malignant tumours are unable to do this owing to the loss of oncogene-induced senescence effectors such as p16^{INK4a} or p53.

Manuel Collado*, **Jesús Gil†**, **Alejo Efeyan***, **Carmen Guerra***, **Alberto J. Schuhmacher***, **Marta Barradas***, **Alberto Benguría‡**, **Angel Zaballos‡**, **Juana M. Flores§**, **Mariano Barbacid***, **David Beach||**, **Manuel Serrano***

*Spanish National Cancer Centre (CNIO), Madrid 28029, Spain

e-mail: mserrano@cnio.es

†MRC Clinical Sciences Centre, Imperial College, Hammersmith, London W12 0NN, UK

‡Spanish National Centre of Biotechnology (CNB-CSIC), Madrid 28049, Spain

§Department of Animal Surgery and Medicine, Complutense University, Madrid 28040, Spain

||Centre for Cutaneous Research, Institute of Cell and Molecular Science, London E1 2AT, UK

1. Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. *Cell* **88**, 593–602 (1997).
2. Lowe, S. W., Cepero, E. & Evan, G. *Nature* **432**, 307–315 (2004).
3. Guerra, C. et al. *Cancer Cell* **4**, 111–120 (2003).
4. Dimri, G. P. et al. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
5. Narita, M. et al. *Cell* **113**, 703–716 (2003).
6. Palmero, I., Pantoja, C. & Serrano, M. *Nature* **395**, 125–126 (1998).
7. Tuveson, D. A. et al. *Cancer Cell* **5**, 375–387 (2004).

Supplementary information accompanies this communication on Nature's website.

Competing financial interests: declared none.
doi:10.1038/436642a

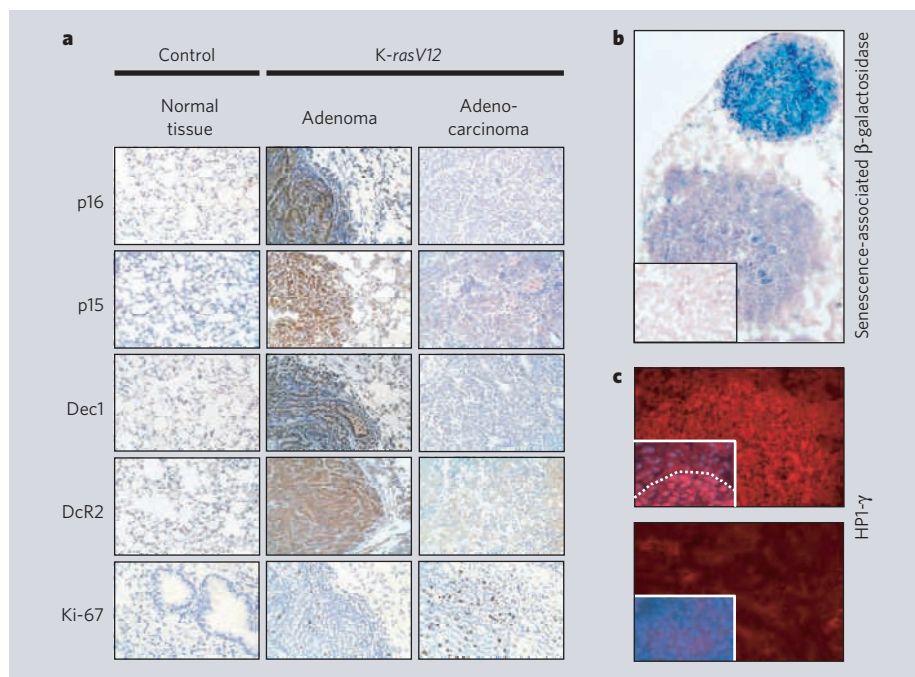


Figure 1 | Premalignant lung adenomas induced by oncogenic *K-ras* are positive for markers of senescence, whereas malignant adenocarcinomas are negative. **a**, Immunohistochemical analysis of the different tissue types for the indicated proteins. **b**, Senescence-associated β -galactosidase expression. A lung section is shown that contains one adenoma (top, blue stained) and one adenocarcinoma (bottom). Inset, negative control. **c**, Immunofluorescence using anti-HP1- γ (red) staining of adenoma (top) and adenocarcinoma (bottom). Insets, double staining with anti-HP1- γ (red) and DAPI (4,6-diamino-2-phenylindole, which stains cell nuclei) (blue). Dotted line, boundary between adenoma and normal tissue. (See supplementary information for further details.)

What Henslow taught Darwin

How a herbarium helped to lay the foundations of evolutionary thinking.

David Kohn, Gina Murrell, John Parker and Mark Whitehorn

The kindly Professor John S. Henslow of Cambridge, well known for arranging Charles Darwin's berth on HMS *Beagle*, was also a rigorous researcher who recorded patterns of variation within and between plant populations and was motivated to understand the nature of species: the big question of natural history as he saw it. The focus of Henslow's research is evident in his herbarium at Cambridge, which holds 3,654 sheets of British plants that he began assembling in 1821. These sheets represent 89% of the species that Henslow recognized in his 1829 *A Catalogue of British Plants*. We have analysed all 10,172 plants on these sheets and infer that he intentionally organized his dried-plant collection to serve as the tool for an inquiry into species and their limits.

Henslow's research on this fundamental question was at its peak during the three consecutive years Darwin attended his lectures (1829–31). Darwin's exposure to his mentor's thinking was part of the exciting intellectual framework that he took with him on the *Beagle*. Indeed, his scientific manuscripts show that direct contact with Henslow provided the context not only for Darwin's botanical studies but also for his comprehension and very acceptance of evolution.

Henslow was a creationist, but with a major difference: he set out to explore the nature of created species as stable entities. Elected professor of botany in 1825, he had already held the chair of mineralogy since 1823. As early as 1821, however, Henslow began establishing a herbarium of British flora. This grew so quickly that by 1829, Darwin's first year as a botany student, Henslow had published the *Catalogue* illustrating his understanding of species.

The herbarium was the product of Henslow's own collecting in Cambridgeshire and Kent, with major contributions from his family, his friends, and most particularly

from the Lancashire solicitor William Wilson. His network included the leading botanists W. J. Hooker of Glasgow and J. H. Balfour of Edinburgh, about 60 collectors strategically deployed to capture floral diversity, and eventually about 30 of his own Cambridge students. One such student was Darwin. On his geological excursion to North Wales with Professor Adam Sedgwick in the summer of 1831 — just before he received the invitation to join the *Beagle* voyage — Darwin collected *Matthiola sinuata* for Henslow. This is the oldest known herbarium specimen collected by Darwin (Fig. 1).

The distinctive feature of Henslow's herbarium was his practice of comparing specimens, which he called 'collation'¹. A collated Henslow sheet carries several plants of a single species from one or more locations, each typically numbered directly on the sheet, with a label recording location, date of collection and collector's name. Collated sheets usually carry two or three plants, but there may be as many as 32. Two-thirds of the sheets are collated and 90% of these show variation in height, leaf shape, branching pattern or flower colour. Collated sheets that show height variation have several distinctive display patterns, such as bell curves and ascending/ descending series (Fig. 2, overleaf). They can depict continuous variation within a single population, or may include plants from across Britain.

Organized approach

Thus Henslow was not just identifying plants: he was organizing his herbarium to emphasize variation within species. Remarkably, he seems to have been the only British botanist at the time doing this. We have surveyed the herbaria of C. C. Babington, J. H. Balfour, William Borrer, W. A. Bromfield, John Downes, R. K. Greville, W. J. Hooker, Leonard Jenyns, W. A. Leighton, N. J. Winch and William Wilson. Henslow's fellow botanists seldom placed more than one plant on a sheet and none practised 'collation'. In Henslow's hands, however, plants received from these same people were collated in a comparative display that illustrated natural variation. This rigorous attention to variation throughout the 1820s was unique to Henslow.

The aim of collation was to analyse the limits of variation within 'created' species. Indeed he was conforming to the orthodox species concept, which included the idea that species only have the capacity to vary within limits^{2,3}. But Henslow recognized that the inherent tension between the stability and variability of species posed a major problem⁴: "Our knowledge...has not been hitherto sufficiently advanced, to



Figure 1 | *Matthiola sinuata*. Herbarium sheet collated by J. S. Henslow from three plants collected by Charles Darwin in 1831 at Barmouth, North Wales, and a single plant collected by Miss Blake at Braunton, Devon. This is the earliest-known herbarium specimen collected by Darwin.

furnish us with any precise rule for distinguishing the exact limits between which any given species of plant may vary." What distinguished Henslow's practice from that of his contemporaries was his intention systematically to turn the creationist species concept into a precise instrument of scientific analysis. This difference of approach may have arisen because Henslow had originally been a physical scientist — a professor of mineralogy who applied the rigour of contemporary crystallography to the species problem⁵:

...[botany] is pretty much in the position which mineralogy occupied before the discovery of the laws of crystallography; mineralogists were frequently in the dark as to what crystals were to be included under one species, and they knew almost nothing of the numerous forms in which any given species might occur... But now, a single crystal at once puts the mineralogist in possession of the primitive form of the species, and he can calculate 'a priori' the possible forms under which it may occur.

Henslow is referring here to the revolution in crystallography articulated by Haüy⁶, who showed in 1801 that complex crystals could be understood as transformations of 'primitive' crystal forms. So when Henslow moved from mineralogy to botany in 1825 he sought a similarly rigorous means to define the natural lines of cleavage within and between plant species. Consequently, in an age of 'splitters' who proliferated new species on the basis of slight differences, Henslow tried to make a precise science out of 'lumping'. He saw varieties where others distinguished species and it was his own collated herbarium that gave him this view. The summary of Henslow's collations was *A Catalogue of British Plants* and the results are most clearly recognized in the 1835 revision, in which he demoted 100 species to the rank of variety. He was thus able to challenge the authority of the great taxonomists of the day: J. E. Smith, A. P. de Candolle, W. J. Hooker and John Lindley.

What Darwin learned from Henslow

What part did Henslow's research play in shaping Darwin's concept of species and his eventual shift to evolution? Darwin learned to read rock formations during his geological tour with Sedgwick in the summer of 1831. Did he also learn to 'read' species from Henslow over the previous three years? We can be sure that Darwin was exposed to Henslow's mode of thinking about species, with its emphasis on discriminating varieties, because this perspective appears in *Principles of*

Descriptive and Physiological Botany (1835), the textbook Henslow based on his lecture course. More importantly, Henslow wove his own research into his teaching. His 1829 *Catalogue* became a set book for his course. In this work, all Cambridgeshire plants were marked for the students in Henslow's classes, who dissected fresh flowers collected on field trips. The *Catalogue* was published in October and so was in preparation when Darwin took Henslow's botany lectures for the first time.

A further attempt to determine the natural lines of cleavage in species came in 1830 — Darwin's second year of botany. Henslow showed that, by manipulating moisture, manuring and shade in garden-grown primulas, he could experimentally reproduce morphological variants observed in the field⁴. Again, the stability of created species is the assumption underlying this work. Henslow supported the linnaean analysis of *Primula veris* with its three varieties: α *officinalis* (cowslip), β *elatior* (oxlip) and γ *acaulis* (primrose) in opposition to J. E. Smith's more modern 'splitting' view. Smith distinguished two separate species: *P. vulgaris* (primrose) and *P. veris* (cowslip) and inclined to the opinion that the oxlip was a hybrid, which he called

P. elatior, that "originated from a Primrose impregnated by a Cowslip"^{7,8}.

Darwin's absorption of Henslow's research activities is seen in his awareness of an immensely important observation that Henslow made during his *Primula* studies — but that he never published. In April 1826, Henslow collected local cowslips and oxlips. He drew whole flowers of each and, more significantly, details of their stamens and ovary (pistil) in half-flower sections. He depicted styles of different lengths in different flowers and also showed the associated differences in anther insertion height. These are what we now refer to as the 'pin' and 'thrum' forms (Fig. 3). Remarkably, he also depicted a form of oxlip flower with short styles and low anthers — the rare short-homostyle form. These drawings, recently discovered in Cambridge, were original observations not known to British botanists.

Stirring memory

Darwin repeated Henslow's *Primula* experiments during the 1850s in the run-up to *On the Origin of Species*, and then, in May 1860, he rediscovered the two flower forms in cowslips⁹. Vaguely remembering that Henslow had seen the same thing three decades before,

he wrote to J. D. Hooker¹⁰: "I have this morning been looking at my experimental Cowslips & I find some plants have all flowers with long stamens & short pistils... others with short stamens & long pistils... This I have somewhere seen noticed, I think by Henslow." But Henslow never published on the lengths of *Primula* pistils, although he did literally notice the different forms. Nowhere in the scores of *Primula* specific-identity articles published between 1830 and 1860 were these different flower forms noticed. Darwin could only have been remembering the cowslip and oxlip drawings that Henslow had shown him when he was a student. Darwin ultimately interpreted the forms of *Primula* flowers as a complex outbreeding mechanism (heterostyly)^{11,12}. But, more importantly, the way he remembered Henslow at the moment he 'discovered' heterostyly demonstrates a hitherto unsuspected familiarity with the heart of Henslow's research.

Much of what Henslow taught would eventually be reflected in Darwin's six botanical books, published between 1862 and 1880. But we see Henslow's core ideas in Darwin's most crucial *Beagle* notes. On the very first Galapagos island he visited, in 1835, while surveying the plants and birds, he asked himself¹³: "I certainly recognize S America in Ornithology, would a botanist?"



Figure 2 | Pattern of display on collated herbarium sheet of *Phleum arenarium*. Eight numbered individuals are arranged in order of increasing height from right to left. Plants 1–5 were collected 3 June 1829 at Mildenhall, Suffolk by J. S. Henslow. Plants 6–8 were collected in June 1822 at Liverpool by W. Wilson.

Darwin went on to collect plants, carefully labelled by island and by date. A decade later, J. D. Hooker used this material to demonstrate what Darwin had already suspected^{14–16}: the high rates of ‘endemism’ (geographically restricted range) in the Galapagos flora. As Sulloway has shown, Darwin could never make the same case stick for his finches, because he had not labelled them by island¹⁷. Indeed, when he first landed in the Galapagos, Darwin obviously thought the plants were more interesting than the birds, so he took due care with labelling. As a faithful Henslow student, he identified his botanical specimens by date and by place.

Moreover, he was collecting plants with a purpose in the Galapagos. This collection was to be a prize for Henslow, who had taught him that oceanic islands tend to be rich in peculiar species, by which he meant endemics. ‘Botanical geography’ was the last topic in Henslow’s course of lectures and in his textbook. Henslow listed several oceanic floras among the 45 botanical regions that he considered to be at least ‘partially examined’^{18,19}; the Galapagos flora is absent from this list. Darwin knew it was important to establish the endemism of his plants. Remarkably, although he still held a creationist view of species, the question of botanical endemism motivated Darwin’s collecting in the Galapagos.

As his exploration of the archipelago proceeded, Darwin recognized an unexpected form of endemism in four of his ornithological specimens. We see a striking development of his thinking in two notes, the first written soon after the *Beagle* sailed from the Galapagos in October 1835 (ref. 20):

This bird which is so closely allied to the Thenca of Chili (Callandra of B. Ayres) is singular for existing as varieties or distinct species in the different Is^{ds}. I have four specimens from as many Is^{ds}. These will be found to be 2 or 3 varieties.

Darwin called the thencas “singular” here because they are different on different islands. But are they different varieties or different species? At this point he settled for the orthodox view — they were “2 or 3 varieties” of the same species. He thus lumped them together, applying the creationist species concept that Henslow had taught him.

Eight months later, in June 1836, while arranging his birds as if they were a Henslow collation, Darwin rewrote these notes on the last leg of the voyage. His opinion had changed²¹:

In each Is^d each kind is exclusively found: habits of all are indistinguishable... When I see these islands in sight of each other, ...tenanted by these birds, but slightly differing in structure and filling the same place in Nature, I must suspect they are only varieties... If there is the slightest



Figure 3 | J. S. Henslow's drawings of variation in style length and stigma insertion height in *Primula*. **a**, Oxlip (8 April 1826, Westhoe) **b**, cowslip (18 April 1826).

foundation for these remarks the zoology of Archipelagoes will be well worth examining: for such facts [would] undermine the stability of Species.

This is the most famous passage that Darwin penned on the entire *Beagle* voyage. Frank Sulloway at the Massachusetts Institute of Technology has established the date of this note and recognizes that Darwin was operating within a creationist species concept on the *Beagle*²². We offer a new interpretation of the passage in the context of what Darwin had learned from Henslow. Darwin now “suspects” they are “only varieties”. He could mean he is now suspicious of the idea that they are merely varieties. Thus they may be species, which would indeed “undermine the stability of Species”. Or he could be going one step further. He could also mean that they are “only varieties” and that is what “would undermine the stability of species” — in which case we are witnessing the birth of Darwin’s most fundamental view, namely that varieties are incipient species.

Either way, here we glimpse Darwin struggling with a radical shift in his Henslowian species concept, just as he crosses the threshold between creation and evolution. Henslow would have preserved species stability, which Darwin no longer attempts to do. But without an appreciation of the depth of his botanical foundation, it has been unclear just how far Darwin was breaking with his past in this passage. We now see that the conceptual framework he received from Henslow was disintegrating and — because of Darwin’s complete facility with the tight logic of that framework — a new way to see species was inevitably crystallizing.

Not surprisingly, Henslow is not acknowledged here by name. We do not cite our teachers for the fundamental ideas they transmit. Rather, they are part of our mental architecture. It seems this was the case with Darwin and Henslow. But in Darwin’s case, the

Henslowian framework he had been given at Cambridge switched into a new configuration. The matter could only be settled by an expert ornithologist, and Darwin no doubt hoped that his carefully labelled Galapagos plants would also provide rich material with which to test the possibility of transmutation.

Henslow had launched Darwin’s voyage when he helped to secure a berth for him on the *Beagle*. But, more significantly, during Darwin’s undergraduate career Henslow had also launched his mind on an intellectual voyage that led from species stability to *On the Origin of Species*.

David Kohn is in the History Department, Drew University, Madison New Jersey 07940, USA. Gina Murrell is in the Cambridge University Herbarium, Department of Plant Sciences, Downing Street, Cambridge CB2 3EA, UK. John Parker is in the Cambridge University Botanic Garden, Bateman Street, Cambridge CB2 1JF, UK. Mark Whitehorn is in Information & Learning Services, University College Worcester, Worcester WR2 6AJ, UK.

- Letter from J. S. Henslow to N. J. Winch, 25 September 1826 (Winch Letters W5.261, Linnean Society of London Library).
- Lyell, C. *Principles of Geology* 1st edn Vol. 2 (London, 1832).
- de Candolle, A. P. *Physiologie Végétale* (Paris, 1832).
- Henslow, J. S. On the specific identity of the primrose, oxlip, cowslip, and polyanthus. *Mag. Nat. Hist. & J. Zool.* **3**, 406–409 (1830).
- Henslow, J. S. On the requisites necessary for the advance of botany. *Mag. Zool. Bot.* **1**, 116–117 (1836).
- Hauy, R.-J. *Traité de Minéralogie* (Paris, 1801).
- Smith, J. E. *English Botany* Vol. 1, 4–5 (London, 1790).
- Smith, J. E. *English Botany* Vol. 8, 513 (London, 1799).
- Darwin, C. R. *Experiment Book* (DAR 157a: 53–57, Cambridge Univ. Library).
- Burkhardt, F. B. et al. *The Correspondence of Charles Darwin* Vol. 8, 191–192 (Cambridge Univ. Press, 1993).
- Darwin, C. R. On the two forms or dimorphic condition in the species of *Primula*, and on their remarkable sexual relations. *J. Proc. Linn. Soc.* **6**, 77–96 (1862).
- Darwin, C. R. *The Different Forms of Flowers on Plants of the Same Species* (London, 1877).
- Darwin, C. R. *Beagle Notebook* 1.17 (MSS Microfilm 532: A13 [1st ser. of notes], Cambridge Univ. Library).
- Darwin, C. R. *Journal of Researches* 2nd edn, 393–397 (London, 1845).
- Hooker, J. D. An enumeration of the plants of the Galapagos Archipelago; with descriptions of those which are new. *Trans. Linn. Soc. Lond.* **20**, 163–233 (1851); (Read 4 March, 6 May and 16 December 1845).
- Hooker, J. D. On the vegetation of the Galapagos Archipelago, as compared with that of some other tropical islands and of the continent of America. *Trans. Linn. Soc. Lond.* **20**, 235–262 (1851).
- Sulloway, F. J. Darwin and his finches: the evolution of a legend. *J. Hist. Biol.* **15**, 1–53 (1982).
- Henslow, J. S. *Sketch of a Course of Lectures on Botany for 1833*, 7 (Cambridge, 1833).
- Henslow, J. S. *The Principles of Descriptive and Physiological Botany* 305–307 (London, 1835).
- DAR 31.2: 341 (Cambridge Univ. Library); Keynes, R. D. (ed.) *Charles Darwin’s Zoology Notes & Specimen Lists from H.M.S. Beagle* 298 (Cambridge Univ. Press, 2000).
- DAR 29.2: 73–74 (Cambridge Univ. Library); Barlow, N. (ed.) Darwin’s ornithological notes. *Bull. Brit. Mus. (Nat. Hist.) Hist. Ser.* **2**, 203–278 (1963).
- Sulloway, F. J. Darwin’s conversion: the *Beagle* voyage and its aftermath. *J. Hist. Biol.* **15**, 325–396 (1982).

Acknowledgements The authors gratefully acknowledge the assistance of Mary Whitehorn, Dale Hwang, Patricia Hawkins and Peter Atkinson.

Correspondence and requests for materials should be addressed to John Parker (jsp25@cam.ac.uk).

REVIEWS

A possible unifying principle for mechanosensation

Ching Kung¹

Of Aristotle's five senses, we know that sight, smell and much of taste are initiated by ligands binding to G-protein-coupled receptors; however, the mechanical sensations of touch and hearing remain without a clear understanding of their molecular basis. Recently, the relevant force-transducing molecules—the mechanosensitive ion channels—have been identified. Such channel proteins purified from bacteria sense forces from the lipid bilayer in the absence of other proteins. Recent evidence has shown that lipids are also intimately involved in opening and closing the mechanosensitive channels of fungal, plant and animal species.

All creatures have mechanical senses: insects hear, and, when touched, worms twitch and sea anemones contract. Touching the front of a unicellular paramecium makes it swim backward; touching its posterior makes it spurt forward. Plant roots and shoots respond to gravity (gravitropism), and stems proportion the growth in their height, versus the growth in their girth, based on the amount of jostling by wind and rain (thigmomorphogenesis). Besides the ear and the skin, animals have many other mechanosensors; for example, the circumventricular organs (for determining systemic osmolarity), baroreceptors (blood pressure), spindle receptors (muscle stretch), proprioceptors (limb positions) and so on. Our bones measure stress during periods of growth or regeneration, and our tongues sense texture and size so that we don't swallow sand or stones.

Mechanical senses differ from other senses. The molecular bases of sensing odorants, hormones, neurotransmitters and other dissolved ligands (solutes) are well understood: the lock-and-key binding of each ligand to the specific binding pocket of its specific receptor on the plasma membrane. Much less is known, however, about the molecules that sense forces such as osmotic force, thirst, touch, vibration and texture. Many membranes are equipped with mechanosensitive (MS) ion channels that respond to turgor in proportion to the surrounding concentration of water (the solvent). Such channel proteins have been cloned and crystallized from bacteria. Examination of these proteins by genetic, electric, chemical and physical means has found that they are able to directly detect and respond to forces from the lipid bilayer. The study of MS channels in plants and animals lags behind, partly because their anatomical complexities resist reductionistic approaches. Nonetheless, recent findings indicate the involvement of lipids in the gating of MS channels of the worm, the fly, the frog oocyte and in mammals. The same types of channel proteins—for example, transient receptor-potential (TRP) proteins—are now found to sense vibration, touch and osmotic membrane stretch. The possibility is emerging that force detection ultimately occurs at the channel–lipid interface. Displacement at the interface, either by deforming the bilayer or by pulling the channel or the bilayer with a tether, provides the energetic drive for the channel to reach the open conformation.

This Review covers many of the principal advances in this field, and draws on widely diverse research projects: from hair cells to bacteria. Traditional compartmentalization of research disciplines,

such as the segregation of micro- and neuro-biology, often hampers dialogue among projects. Yet, life's basic machineries—such as DNA replication, transcription, translation, the Krebs cycle, electron transport, and now ion filtration and voltage gating—are found to be universal, even though they were originally based, in large measures, on experiments done in microbial systems. This Review asks whether there is a common physicochemical basis that determines how channel proteins sense force, which may serve to unite the varied biological manifestations. Even though touch, hearing and osmosis are seemingly disparate fields of research, they all deal with a single physical parameter—force. Regardless of frequency or duration, a dyne is a dyne.

MS channels allow bacteria to withstand rain

Life is largely aqueous chemistry—we went to Mars looking for evidence of water and, by inference, life. Cells comprise ~80% water, and de- or over-hydration can be lethal. Osmotic force, a measure of water content, is therefore fundamental to cell survival. Even though water is 55.6 M, it only requires a 10 milli-osmolar difference across a barrier to produce an osmotic pressure of ~180 mm Hg ($\sim 2.5 \times 10^5 \text{ dyn cm}^{-2}$), which generates a $\sim 12 \text{ dyn cm}^{-1}$ stretch force on the surface of a 2- μm -diameter sphere. Surprisingly, the proteins in bacteria that directly gauge such stretch forces were only discovered in the last decade, and only by serendipity. When a cell is caught in the rain, or following laboratory dilution, the inward diffusion of water through the lipid bilayer—water channels are not needed—generates a huge turgor that can rise to hundreds of atmospheres (10^8 dyn cm^{-2}): far beyond what the cell envelope can withstand. It is known, from work first carried out in the 1950s, that *Escherichia coli* bacteria release their osmolytes (solutes) upon an osmotic down-shock (Fig. 1a). For instance, a 1-in-100 aqueous dilution of the culture drains the cellular ^{14}C -labelled proline pool by >95% (ref. 1). Tracing other osmolytes, such as K^+ , lactose and ATP, gave similar results². However, osmolyte-depleted bacteria retain their macromolecules, and begin protein synthesis within minutes upon return to normal medium¹. The identity of the 'emergency valves' for solute release remained elusive for almost half a century. Though MscL and MscS (mechanosensitive channel of large or small unitary conductance) were postulated to be these valves when their electrical activities were discovered in 1987 (ref. 3), this role was not proved until 1999 when Booth and co-workers⁴ showed that ΔmscL

¹Laboratory of Molecular Biology and Department of Genetics, University of Wisconsin, 1525 Linden Drive, Madison, Wisconsin 53706, USA.

$\Delta mscS$ double mutants lyse even upon a rather mild down-shock. MscL and MscS have redundant functions and single mutants have no clear phenotype, so they were not discovered in the extensive and intensive genetic dissection of *E. coli* physiology in the last century, but were, in fact, identified in a patch-clamp survey (discussed below). MS channels are just one—albeit the quickest one—of a bacterium's many defences against de- or over-hydration. The slower types of these defences are those that are transcriptionally controlled.

Prokaryotic MS channels

MscL and MscS were encountered in the first electrophysiological survey of bacterial membranes, conducted by Martinac *et al.*³. Voltage-controlled (-clamped) patches of *E. coli* membranes produced giant steps in unitary current when the patches were subjected to suction³, or when the solution bathing the patch was diluted⁵. These MS-channel activities remain following reconstitution into artificial liposomes after membrane dissolution or protein purification⁶ (Fig. 1b). The non-selective unitary conductances of MscL (~ 3 nS) and MscS (~ 1 nS) are 10–1,000-fold greater than those of the more selective channels commonly studied. Such large signals and unlimited bacterial material allowed the tracing of the channel

activity in chromatographic fractions to a single protein, which, in turn, led to the discovery of the *mscL* gene^{7,8}. The corresponding small protein has two transmembrane (TM) helices, M1 and M2 (refs 7, 9; Fig. 2, left). Chang *et al.* resolved the crystal structure of the *Mycobacterium tuberculosis* MscL homologue as a homopentamer with the five M1 domains converging to close the pore at the cytoplasmic side¹⁰ (Fig. 2, centre). In 2001, computational modelling and cross-linking experiments led Sukharev *et al.*¹¹ to a model in which all TM helices recline and rotate, like the iris of a camera, to open a large pore of some 30 Å in diameter (Fig. 2, right). The following year, through observations with electron paramagnetic resonance spectroscopy after site-directed probe attachment, Martinac and co-workers^{12,13} concurred with the main premise of this model. The cloning⁴ and crystallization¹⁴ of the *E. coli* MscS protein showed that it is a homoheptamer of three-TM subunits with seven converging M3 domains: entirely different from MscL. Genetic, biochemical, biophysical, simulation and other studies on MscL and MscS are periodically reviewed^{15–18}.

Forces from lipids gate prokaryotic MS channels. Exercising ultimate reductionism, mechanosensitivity was found to be intact when pure MscL protein was placed in bilayers of one^{12,13} or two⁷ defined lipids (Fig. 1b). Here, the stretch force detected by the protein must come from the lipids themselves, since there are no other components that could contribute. MscL follows a Boltzmann distribution where the mechanical energy partitions the MscL protein between its closed and open conformations, with the mid-point tension (50% of the channel being open) at ~ 12 dyn cm⁻¹ (ref. 19)—a sensitivity presumably tuned for its biological role (Fig. 1a). (The threshold tension of MscL is much lower; wild-type MscS (ref. 20) and MscL gain-of-function mutants^{21,22} have even lower thresholds.) MscL functions in bilayers made of ordinary lipids with charged or uncharged heads, saturated or unsaturated tails, and in various mixtures. Shortening the length of the fatty acid chain from 20, to 18, to 16 carbons reduces the energy barrier between the closed and the open state, but does not trigger spontaneous channel opening^{12,13}.

The prevailing model of the behaviour of MscL and MscS addresses the forces within the lipid bilayer itself. Unlike the aqueous solution, the bilayer is highly anisotropic: having very different physical properties at different depths. The free-energy reduction in ordering waters and lipids at the interface is reflected in a large surface tension between the lipids' polar head groups and the non-polar tails. However, pressures nearby serve to balance this tension, allowing the bilayer to self-assemble into a stable structure. The force profile of pure lipid bilayers has been calculated by Cantor^{23,24} (Fig. 3a) and examined with molecular-dynamics simulation^{25,26}. (How the protein–lipid interaction may distort the profile at the interface is unclear, though it is being analysed²⁷.) The amplitudes of these forces are in the order of hundreds of dyne cm⁻¹ (ref. 25); much larger than the lytic tension (tens of dyne cm⁻¹) of the bilayer. Any protein embedded in the bilayer is subject to these strong, localized push or pull forces. Altering the force profile by membrane stretch, or by channel or lipid displacement through a tether (discussed below), may make it energetically more favourable for the channel protein to assume a new conformation; for example, the open state (Fig. 3b). Among the molecular-dynamics simulations carried out by several groups, Gullingsrud and Schulten²⁶ directed forces to residues within the MscL protein at the level of the lipid's glycerol backbone between the head groups and the tails, where tension is maximal (yellow arrows, Fig. 2). Such a simulation can reveal steric clashes or structural disintegration, which should not happen when a channel opens. They traced the positions of the 111,079 atoms of 1 MscL protein, 365 lipids and 22,308 waters, and found that MscL indeed opens on a 10-ns timescale in an iris-like manner, similar to the original model^{1–13}.

Besides external forces, the composition of the lipid bilayer itself can affect its internal forces. Adding chemically unrelated cationic

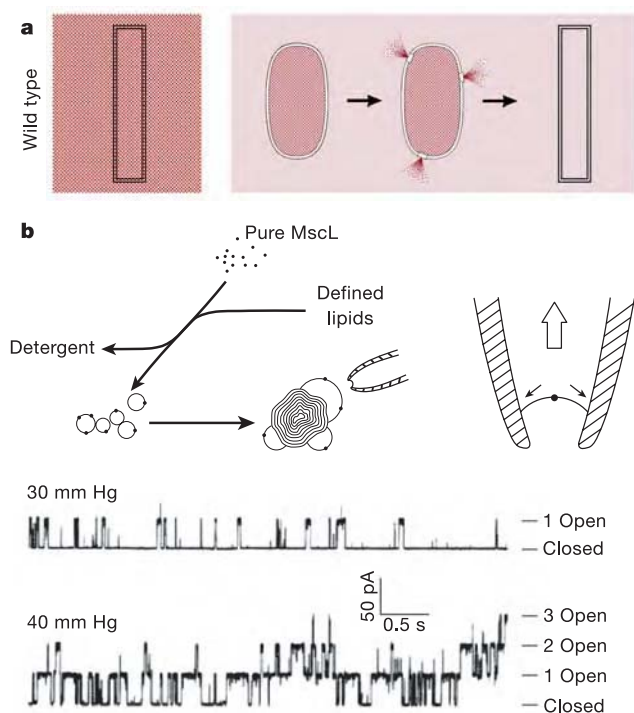


Figure 1 | Bacterial channels function as emergency release valves *in vivo*, and the mechanosensitivity of pure MscL channel protein *in vitro*. **a**, An *E. coli* cell in a normal environment (left) and in the rain (or upon dilution in the laboratory, right). A bacterium (shown as a rod), having adjusted its cytoplasm to the relatively high osmolarity of the surrounding milieu (shown in dark red, the red dots being solutes, not water), is confronted with a sudden dilution of its environment upon the onset of rain (light red). Entry of water (not shown) through the lipid bilayer swells the bacterium (now oval-shaped) and stretches open the MS channels to jettison solutes (red puffs), enabling it to reach a new equilibrium and escaping osmolysis (and returns to being rod-shaped). **b**, Purified MscL protein is reconstituted into multilamellar liposomes after replacing the detergent with lipids. Induced liposome blisters can be sampled with a patch-clamp pipette. A suction applied to the pipette (broad open arrow) creates tension (small filled arrows) in the membrane patch and activates MscL proteins. The increase in the number of channel openings in a patch (shown as unitary-conductance steps at the marked levels) is evident when the suction applied to such a patch of lipid bilayer increases from 30 mm Hg to 40 mm Hg (4×10^4 dyn cm⁻² to 5.3×10^4 dyn cm⁻²) (modified from ref. 15).

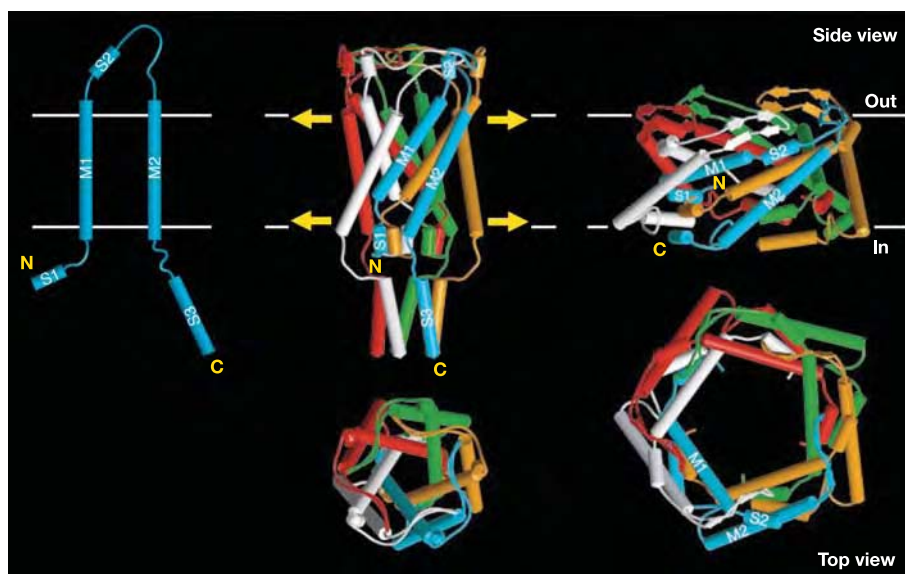


Figure 2 | Opening MscL in *E. coli*. Helical segments (S1, S2, S3) and transmembrane helices (M1, M2) in one MscL subunit, as deduced from sequence⁷ and other analyses⁹ (left). Side (upper centre) and top (lower centre) views of the closed channel backbone structure of the *E. coli* MscL protein, by analogy to the crystal structure of the *M. tuberculosis* MscL homologue¹⁰. The open structure deduced from both modelling and experimentation¹¹ (right). Unlike MthK, the prokaryotic K⁺ channel that is

equipped with a second constriction (the K⁺ filter), MscL is like the acetylcholine receptor/channel, in which the open gate doubles as the filter. Here the opening is huge (~30 Å in diameter): befitting its ability to release solutes indiscriminately (as shown in Fig. 1a). The work to increase the area under tension constitutes the free energy difference that partitions the open and closed states. (Modified from ref. 11.)

amphipaths to the membrane of a red blood cell causes it to form cups, while adding anionic amphipaths causes it to form bulges (to crenate)²⁸. Regardless of the cause, changes in the bilayer's geometry will distort the force profile contained within it. Indeed, these agents have been shown to reversibly activate MscL and MscS. The potency of the amphipaths to activate such channels is proportional to their lipid solubility²⁹, and they are effective only when added to one monolayer but not both¹³. The shapes of the added lipids are important, as evident from the behaviour of gramicidin A upon bilayer modifications uncovered by Andersen and co-workers³⁰. The usual bilayer-forming phospholipids can be approximated as rods (Fig. 4, shown in red), and the micelle-forming lysophospholipids (with a single fatty-acid chain) as cones (Fig. 4, blue). Polyunsaturated fatty acids (PUFAs), such as arachidonic acid (AA, a precursor of prostaglandins), can be regarded as inverted cones with smaller heads than tails (Fig. 4, green). Cones, or inverted cones, differentially entered into one monolayer can change the local curvature and the internal force profile, redistributing the tension between the two leaflets (Fig. 3a). Indeed, the addition of lysophosphatidylcholine triggers the opening of MscS (ref. 13). Structurally diverse anaesthetics, which are all lipid soluble, have been theorized to change the bilayer force profile³¹. Indeed, experiments have shown that procaine and tetracaine can activate MscS (ref. 29).

Eukaryotic MS channels

Plants (for example, *Arabidopsis thaliana*) have clear homologues of MscS. Most animal cell membranes present MS conductances, but only a few have been traced to known gene products. Although these differ from MscS and MscL in sequence, some of their properties are quite similar. Patel and co-workers found that the polymodal K⁺ channel of mammals, TREK-1 (two-pore domain weak inward-rectifying (TWIK)-related K⁺ channel), can be activated by both force and osmolarity. Similar to MscS and MscL, it is also activated by the bulge-forming amphipathic chemicals such as trinitrophenol (called crenators), but inhibited by the cup-forming amphipaths such as chlorpromazine^{32,33}. The cone-shaped lysophosphatidylcholine activates it, and the exaggerated cone lysophosphatidylcholine activates even more effectively^{32,34}. Structurally diverse anaesthetics

such as chloroform, halothane, isoflurane and diethyl-ether also activate TREK-1 (ref. 35). Negatively charged lipids such as PIP₂ (phosphatidylinositol 4,5-bisphosphate) or phosphatidic acid, when presented to the membrane inner leaflet, also activate TREK-1 (ref. 36), and the charged cone lysophosphatidic acid strongly activates it³⁷.

MS and some other types of channels are inhibited by the small lanthanide Gd³⁺ (refs 38, 39). The mechanisms of this inhibition are complex and include its interaction with the lipid bilayer^{39,40}. Sachs and others found an amphipathic 34 L-amino-acid peptide⁴¹ in a tarantula venom to inhibit the MS conductance in cultured mammalian cells⁴². They also found the synthetic enantiomer of 34 D-amino-acids to be just as effective⁴³, and concluded that the peptide could not have bound to a channel protein lock-and-key-like, but entered the bilayer to affect the channel's surrounding environment.

Eukaryotic cells have extensive cytoskeletons near the bilayer, which are often assumed to be the transmitters of force. One needs to examine this assumption carefully. Hamill *et al.* examined the MS channels in the complex cell surface of *Xenopus* oocytes⁴⁴. From this surface, they induced blebs that had little or no cytoskeletal elements and continued to encounter MS-channel activities in the bleb membranes^{16,44}, and it was from such membranes they now traced the channel activities to TRPC1 (transient receptor-potential canonical 1; ref. 45; discussed below). The cortical cytoskeleton network often folds a large excess of membrane bilayer into microvilli or caveolae. This network is far more extendable than the bilayer itself, allowing cells to swell without increasing the total bilayer area or tension, which explains why MS currents sometimes cannot be found in whole cells but only in excised patches, where the cytoskeleton is lost^{46,47}.

Animal sensory cells often have organized microtubules such as the ciliary axoneme (see below). There are also specialized microtubular processes along the length of the long touch-sensing cells in *Caenorhabditis elegans* studied by Chalfie and co-workers^{48,49}. The touch-insensitive (loss-of-function) or touch-cell degeneration (gain-of-function)^{48–50} worm phenotypes led elegantly to the identification of a series of mechanosensitivity genes (the *mec* genes)

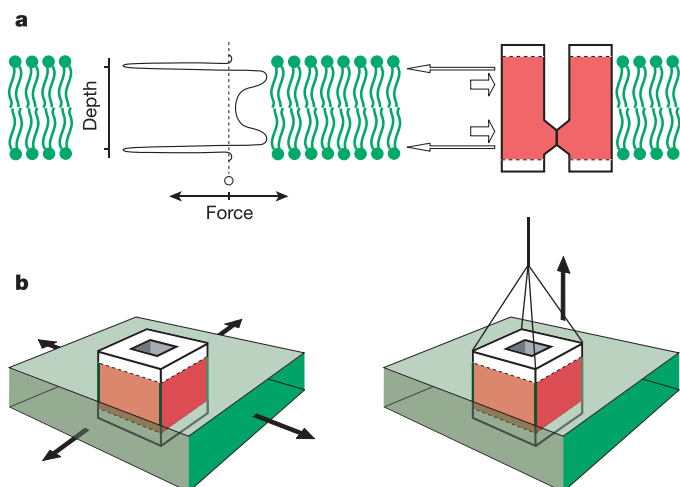


Figure 3 | The intrinsic forces in the lipid bilayer, and how applied forces can open MS channels. **a**, The intrinsic force profile plotted as its direction and magnitude along the depth of the bilayer (left)²³, and a cartoon of a channel protein in section (right), showing how the sharp tension (narrow arrows) near the lipid necks balanced by more diffused pressure nearby (broad arrows) is exerted on the channel–lipid interface (red). **b**, The forces at the crucial channel–lipid interface (red) will change when the bilayer (green) is stretched or bent (left), or when the channel is displaced from the bilayer through a tether like an elevator (right). It is also possible that the tether, through ancillary proteins, pulls on the lipids surrounding the channel (not shown). In all cases, changes in the force profile at the interface (red) can become the ultimate trigger for the channel's conformational change.

including *mec-4* and *mec-10*. These genes encode channel subunits akin to the epithelial Na^+ channels, and conduct channel currents in frog oocytes^{51,52}. The channels, and their associated components, form punctate clusters that are evenly spaced along the worm's microtubular processes. Among the MECs are an extracellular matrix protein (MEC-1) and special tubulins (MEC-7, MEC-12). Previous models described a trans-cellular complex, similar to the original trapdoor model for the vertebrate hair cell (now modified, see below), in which the displacement of the matrix, resisted by the microtubules, gates the channels in between^{48,49}. Recent analyses showed that MEC-1, and at least two other gene products in the matrix, are needed to form the punctate clusters. However, removal of the special microtubules through mutation have only limited or no effect on the structure or the function of these clusters^{53,54}. These observations, and the fact that the transduction current is reduced but not abolished in *mec-7* mutants⁵⁵, have now questioned whether the MS channel is directly gated by tethering to microtubules⁵³.

Tethering to rigid elements, even if true, does not necessarily imply force transmission. Current cell biology teaches us that meaningful protein–protein contacts, whether ephemeral or long-lasting, are the norm. For example, the signalling machinery of the *Drosophila* photoreceptor, not known for its mechanosensitivity properties, is comprised of rhodopsin, G protein, enzymes and channels, all tied together into a “transducosome”⁵⁶ or “signalplex”⁵⁷ tethered to an actin cytoskeleton, which serves to deploy and station the complex near the cell surface.

What trips TRP channels?

Several MS channels making recent news belong to the TRP superfamily^{45,58,59}, the founding member of which was identified in a near-blind *Drosophila* with a transient receptor-potential in its electroretinogram⁶⁰. Forward genetics starting from mechanosensitive phenotypes led us to these TRP channels without preconceived bias. This has happened six times recently—in the worm^{61,62}, fly^{63,64}, mouse⁶⁵ and human⁶⁶—and cannot be a coinci-

dence. At least seven homologues of the TRPs identified in this way then became candidates in the ensuing investigations and have now also been tied to mechanosensation.

TRPV4 (transient receptor-potential vanilloid-4; previously called the vanilloid receptor-related osmotically activated channel, or VR-OAC), the most studied mammalian MS TRP channel, is found in many tissues including the circumventricular organs of the central nervous system and inner-ear hair cells. The heterologously expressed whole-cell current through TRPV4 can be activated by cell inflation with pressure⁶⁷, by mechanical shear force from bath flow⁶⁸, or by mild hypo-osmotic challenges^{69,70}. Deleting the three amino-terminal ankyrin repeats does not significantly impair TRPV4's response to hypo-osmolarity⁶⁹. The Bargmann laboratory found the first MS TRPV through *osm-9*-mutant worms, which fail to recoil from an osmolar solution or a nose touch⁷¹. Worms with normal OSM-9 channels, but deficient in the synthesis of a set of 20-carbon PUFAs (including AA, Fig. 4), showed similar deficits in behaviour that could be restored by a dietary supplement of PUFAs. One such PUFA acts as an irritant that appears to directly activate the channel⁷². The rat TRPV4 is only 24% identical to the worm's OSM-9. Nonetheless, Liedtke *et al.* found that a *trpv4* transgene complements the defects of the *osm-9*-mutant worm, and that the restored behaviours have a threshold and temperature optimum of a warm-blooded animal⁷³. Furthermore, a version of TRPV4 deleted of both its amino- and carboxy-terminal cytoplasmic domains, the presumed cytoskeleton-binding sites, remained able to complement the phenotype⁷³. The single-channel current of TRPV4 has received far less scrutiny. Unitary conductances of 310 pS (ref. 69), 60 pS (ref. 74), or 30 and 88 pS (ref. 70) have been variously reported from similar experiments. The only report on attempts to activate TRPV4 unitary conductance by direct patch suction was negative⁷⁰. In contrast, a 20 pS MS conductance (in frog Ringer's solution) from oocyte membranes has recently been traced to TRPC1 through liposome reconstitution by Maroto *et al.*⁴⁵, who also showed that heterologously expressed human TRPC1 produces MS unitary currents. Although reconstitution of purified TRPC1 has not yet been reported, this work comes closest to showing that certain TRP channels receive their gating force directly from the lipids.

Channels associated with mechanosensation appear in nearly all TRP subfamilies: TRPV, TRPC, TRPA (ankyrin-like), TRPP (polycystin), TRPN (NOMPC, or no mechanoreceptor potential C), TRPY (yeast) and probably TRPML (mucolipins), each having rather different cytoplasmic domains, suggesting that these domains are not

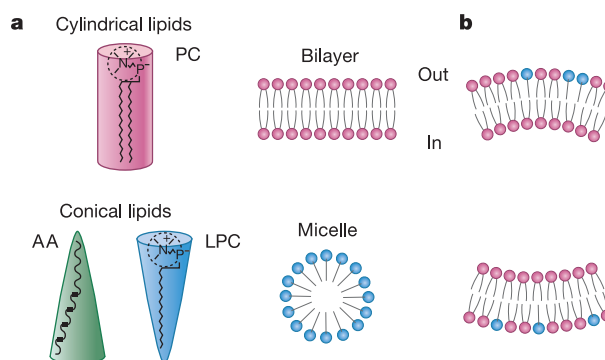


Figure 4 | The shape of bilayer components affects its geometry and intrinsic forces. **a**, Bilayer-forming phospholipids (shown in red), such as phosphatidylcholine (PC), can be approximated as rods. Micelle-forming lysophospholipids (blue), such as lysophosphatidylcholine (LPC) with only one fatty acid chain, can be regarded as cones. Polyunsaturated fatty acids (green), such as arachidonic acid (AA), approximate inverted cones. **b**, The differential addition of cone-shaped lipids (or other amphipaths, not shown) into one of the monolayers can alter the shape, and therefore the intrinsic forces, of the bilayer. (Modified from ref. 32.)

the crux of mechanosensitivity. TRP channels are polymodal. Heterologously expressed TRPV4, for example, can also be activated by heat, by a phorbol ester, by anandamide and by AA (Fig. 4). The well-known heat-sensing vanilloid receptor, TRPV1, identified in the Julius laboratory⁷⁵, is also activated by low pH and endogenous inflammatory ligands. The *trpv1* knockout mice and their bladder urothelium have abnormal response to hypo-osmolarity⁷⁶. Though a universal switch for different mode of stimuli may be appealing, there is evidence that different stimuli use different pathways to open TRPV4 (ref. 77).

Strings attached—the roles of tethers

Several TRP channels have now been located within some of the complex MS organs. For example, in the fly chordotonal organ, a matrix material presses against the dendritic cap on top of the sensory cilium, which swells into a dilation about one-third of the way from the tip. Kim, Kernan, and their co-workers found that the only two TRPV channels encoded within the *Drosophila* genome, NAN (encoded by the *Nanchung* gene)⁷⁸ and IAV (encoded by the *Inactive* gene)⁷⁹, apparently form heteromeric channels that transduce vibrations into receptor potentials. NAN and IAV proteins are found on the cilia but not the rest of the neuron, and it is restricted at and below the ciliary dilation, some distance from the tip (Fig. 5a). How the movement of the cap-cell matrix translates into the forces experienced by the IAV-NAN channel in the ciliary membrane that lies below it is unclear. The NAN (ref. 78) and IAV (ref. 79) proteins have been individually expressed in cultured cells, where they confer hypo-osmotically induced whole-cell current and a rise in cytoplasmic Ca^{2+} levels. A parsimonious interpretation would be that the channel experiences the vibration as a stretch along the membrane plane. Whether there are other proteins that tie them to the axoneme, whether such tethers transmit force, and whether there are roles for other channel subunits (for example, NOMPA, or no mechanoreceptor potential A, located at the distal end of the cilium)⁷⁹ await clarification.

The vertebrate hair cell is a clear case in which the gating force is passed on to the MS channel through a tether; though it is debatable whether the tether directly pulls a certain domain from the rest of the channel protein, which is held in place by a resistive force, as originally proposed. Molecular identifications recently enabled great strides in this system. First, TRPN was found in sensory hair cells, and to be required for hearing and balance in zebrafish⁸⁰. Then, cadherin 23 was discovered to be a major component of the tip link^{81,82}, and judged to be too stiff to comprise the gating spring. Recently, Corey *et al.*⁸³ showed that *TRPA1* messenger RNA appears at the appropriate time during hair-cell development, and that *TRPA1*-expression knock-down curtails transduction *in vivo*. *TRPA1* protein is found at the upper part of the stereocilia, though not just at the very tip. It is also located in the pericuticular zone (perhaps for secretion) and, when present, the entire kinocilium (the true cilium with a microtubular axoneme) (Fig. 5b). With such rapid progress, we are looking forward to the identification of other gene products in this system in the very near future. Meanwhile, fleshing out the biophysical scheme with the molecules so far identified has already led to modifications of the familiar 'trapdoor' model of hair-cell mechano-transduction (Fig. 5c, left). Calculation and simulation show that the long N-terminal ankyrin repeats of *TRPA1* are compatible with the elastic property of the gating spring^{83,84} (Fig. 5c, right). (The cartoons in Fig. 5c, and similar representations elsewhere, should not be taken literally as the site(s) and the nature of string attachment, in addition to the number and identity of channel subunits and other elements, are unknown. The upper tether can be attached to the gating springs beneath the membrane instead of the channel body.) In the more thoroughly examined multimeric channels, such as MthK (the K^+ channel of *Methanobacterium thermoautotrophicum*; ref. 85), MscL (ref. 12) and MscS (ref. 14), the channel gates open like the iris of a camera. It is not yet clear how

the one-dimensional movement of a stereocilium leads to the two-dimensional opening of the TRP tetramers on the side of the cilium, and whether or how the ciliary membrane may move during transduction. *TRPA1* is also expressed in neurons of the dorsal root ganglia, the trigeminal nerve, and photoreceptors, making one wonder for which functions the ankyrin repeats serve in these other locations. A role of ankyrin in assembling a TRP tetramer has been suggested⁸⁶.

Returning to the main theme of this Review, the vertical movement of a channel by the tip link does not necessarily negate the involvement of the lipid bilayer. It seems possible that such 'elevator'-like movement would still eventually displace the channel with respect to the bilayer's intrinsic force profile (Fig. 3b, right). The mismatch and asymmetry produced by the displacement can, in fact,

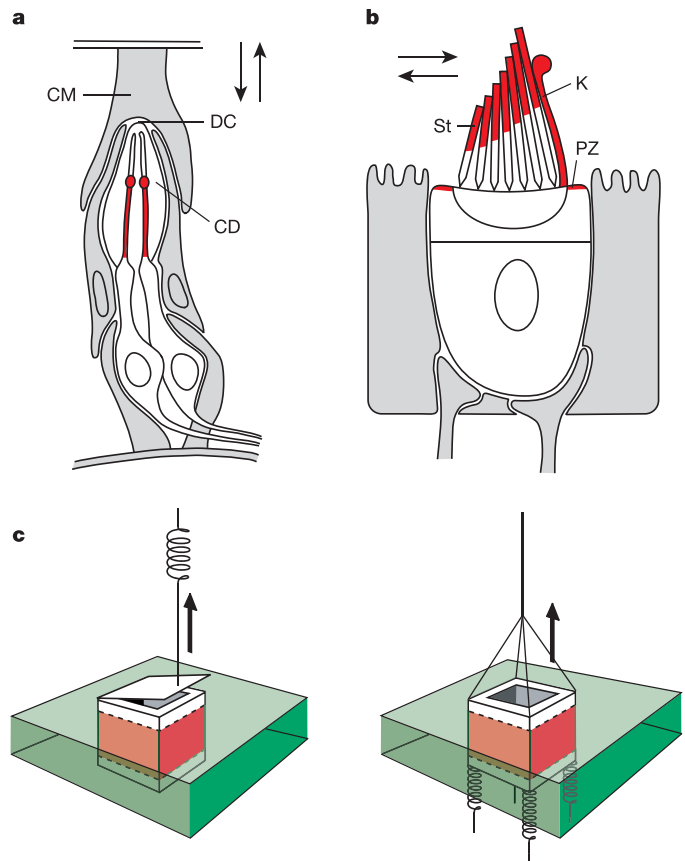


Figure 5 | TRP channels in auditory sensory cells. TRP channels have been located in complex auditory sensory cells, even though the mechanism by which ciliary vibrations (arrow pairs) lead to the iris-like opening of the channels on the side of the cilia is not clear. **a**, The antennal chordotonal organ of *Drosophila*. CM, cap-cell matrix; DC, dendritic cap; CD, ciliary dilation. Red marks the location of NAN (a TRPV-type channel subunit encoded by the *Nanchung* gene). (Redrawn from ref. 78.) **b**, A vertebrate hair cell. St, stereocilia; K, kinocilium; PZ, pericuticular zone. Red marks the location of *TRPA1*. (Modified from ref. 83.) **c**, Models of the vertebrate hair-cell transduction channel. Molecular identifications have transformed the biophysical trapdoor model (left) to one with a *TRPA* channel and a stiff cadherin-containing tip link (right). The elastic element of transduction is now assigned to the ankyrin repeats in the four (presumably) *TRPA* subunits^{83,99} (shown as coils), which are presumed to be attached to cytoskeleton and/or myosin (not shown). This current model is compatible with one in which the displacement of the channel protein, with respect to the lipid bilayer, ultimately triggers the channel conformation change as shown in Fig. 3b, right. However, none of these models should be taken literally since we do not yet know the true composition of the transduction channel(s) and how the various channel components contact each other and the lipids. See the main text for some possible variations.

be the final mechano-energetic trigger for the required channel-configuration change. It is even possible that the tether pulls on the rim of an elastic carrel, say, and passes the force through the lipids that surround the channel (Fig. 3b, left), but this possibility has not been investigated. The 'trapdoor' (Fig. 5c, left) is usually interpreted as the separation of individual protein domain(s) from the rest of the channel protein by mechanical work¹⁷. The 'elevator' entails displacing the entire channel protein from its normal lipid environments, generating tension around the entire circumference, and leading to an iris-like opening (Fig. 3b, right). Either the trapdoor or the elevator model can accommodate the mechanical resistive and elastic elements traditionally described.

It is difficult to imagine TRPA1 being indifferent to the lipids, given its activation by lipid-like compounds. Gillespie and co-workers showed that PIP₂ localizes towards the tip of the hair bundle and is required for both the mechanical transduction current and its adaptations⁸⁷. The mouse TRPA1 was first reported to be a sensor of noxious cold⁸⁸. It is also activated by bradykinin, and the oils of mustard, cinnamon, wintergreen and the like, as well as cannabinoids^{89,90}.

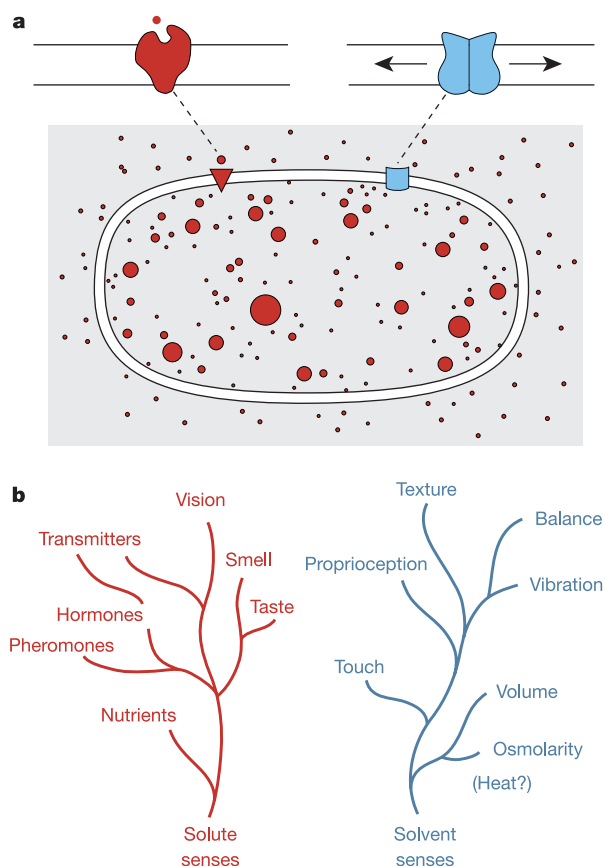


Figure 6 | The disparate sensing of solutes and solvent. **a**, A diagram of an imaginary early cell equipped with two types of receptors that are required to sense solutes and solvents—the two ingredients of life's chemistry. The dots in the grey background represent water molecules (the solvent) and the red circles represent solutes (molecules dissolved in water). When a cell accumulates solutes, the internal water concentration is reduced and the tendency of water to enter the cell results in a turgor. Both the lock-and-key type of receptors (red) for different solutes (ligands), as well as the turgor sensors (blue) for water (the solvent), are needed for even an early cell to survive. **b**, A hypothetical diagram (not to be mistaken for phylogenetic trees) on the grouping of various senses that emphasizes the discrete separations of the lock-and-key type of sensing of the solutes (red) from the force-from-bilayer type of sensing of the solvent (blue). A further description can be found in the text. (Modified from ref. 100.)

Solute senses versus solvent senses

Specialized sensory cilia develop from embryonic primary cilia, and have evolved from motile cilia similar to the ones still found in *Paramecium*, *Chlamydomonas* and so on. Can we trace the origin of TRP-based mechanosensation beyond motile protists? An MS TRP channel is found in the vacuolar membrane of yeast⁹¹, where it detects osmotic forces *in vitro*⁹² and *in vivo*^{93,94}. Because all cells have to deal with osmotic force, it may hold a key to the evolutionary origin(s) of mechanosensation.

MscL and MscS proteins are found in most free-living *Bacteria* and *Archaea*. Thus, the principle of mechanical gating by forces from the lipid bilayer most likely evolved before the divergence of these two domains of life some 3.5 billion years ago. It makes teleological sense for these devices to have evolved early on. When early cells separate two solutions, and horde solutes into one, a water gradient—and therefore a turgor—must develop on the chemiosmotic partition. This turgor has apparently been used ever since, as all extant cells today still have to be turgid to be in a growing steady state. This turgidity helps to break the bonds in the network of rigid elements (membrane, cell wall, extracellular matrix, cytoskeleton and so forth) so that new material can be inserted. Off steady states—the sudden large rise in turgor at the onset of rain (over-hydration) and the large fall in turgor in prolonged hot sun (dehydration)—are likely to have exerted selective pressure on early cells to evolve mechanosensors such as the ancestors of TRP, MscS and MscL. Once the basic principle of activation by lipid forces is employed, it seems unlikely that nature would abandon the principle in the detection of other forces later on in evolution, even as newer generations of protein types out-compete the old. As reviewed above, many extant TRP channels in animals still seem to respond to osmotic forces exerted on the lipid bilayer even though their relatives are specialized for hearing, balance, touch, or texture.

Today, we find a myriad of surface receptors for irritants, odorants, hormones, neurotransmitters and growth factors on cell surfaces. They are not sequence homologues, but they could have originated from one or a few receptors by divergent, as well as convergent, evolution. More importantly, the same physicochemical principle of lock-and-key fitting underlies all ligand sensing. Yet one cannot imagine a protein with a lock-and-key water-binding site that discriminates the milli-osmolar differences that cells do, while water exists in tens of molar. Rapid changes in water concentration must therefore be sensed with a different mechanism. From this perspective, it is not difficult to imagine that a parallel myriad of mechanosensors may have evolved from some simple osmotic-force-sensing device in early membranes⁹⁵. Figure 6a shows an 'ur-cell' with the two types of receptors for the various solutes (for example, nutrients and wastes) and for the one solvent (water). Figure 6b shows a schematic representation of how different senses might have evolved throughout the 3.5 billion years of evolution. This is not a molecular dendrogram. The branching pattern has no precise meaning except that the senses were few in number in the beginning, and that sensing of nutrients and water is fundamental, ancient and disparate. This diagram emphasizes the distinctness of the two classes of sensations that originated to deal with solutes versus solvent—the two basic ingredients of life's chemistry.

Complexities and unknowns

The disparity between solute sensing and solvent sensing can help clarify our thinking. Many solute (ligand) receptors do not require a membrane. It is also 'crystal clear' (from X-ray diffraction studies) that some MS channels have no specific ligand-binding sites (for example, MscL, MscS). These examples should not lead us to assert that no receptors employ both principles. Given nature's propensity to tinker opportunistically, it is even likely for some receptor proteins to evolve specific ligand-binding pockets that face the two aqueous compartments, as well as a lipid-facing surface that transduces forces. This complexity may explain certain polymodality⁷⁴ or

other differences (for example, the TRPV-dependent behaviour of live worms requires certain features of PUFAs but not necessarily AA⁷², while heterologously expressed TRPV4 *in vitro* seems to require an AA metabolite⁷⁴). Complexities can also arise from possible 'signalling lipids' that may act at the protein–lipid–water junction. For example, adding negative charges to the inner leaflet with PIP₂ (ref. 87) will have local electrostatic effects^{36,37} as well as secondary effects on lipid distribution and bilayer force profile.

In *On the Soul* (ref. 96), Aristotle argued that there could be no more than five senses, and folded the senses of hot and cold into touch. He might have been on to something. The detection of heat and that of impact seem to convolve in biology. Heat-induced bilayer rearrangement may alter the membrane tension that gates the polymodal TRP channels⁹⁷. Teleologically, have they been designed to sum heat and force (for example, into 'pain')? Mechanistically, can we separate the force sensor from the heat sensor within the same protein? The truism that everything is mechanical, and heat-sensitive, is not helpful. How can a protein–lipid–water complex be designed to be especially sensitive to a rise in thermal agitation? And how can a similar complex become sensitive to its fall? How can some TRPs be constructed to have an unusually high Q₁₀ value (in excess of 10)? If the thermosensor is coupled to a voltage sensor⁹⁸, what is the coupling mechanism? Doesn't the temperature sensitivity of a channel–lipid complex depend on the entropy of its hydrophobic interactions with water? Why are the temperature mimics (such as capsaicin, menthol and ginger) oily? Why are the agents that numb the sensation amphipathic? Why is the potency of different general anaesthetics proportional to their solubility in olive oil (the Meyer–Overton rule)? Could the protein–lipid–water junction hold even more secrets? Much work lies ahead if the answers to these questions are to be found; perhaps, just like after a long spell of rain, the floodgates of knowledge are about to be opened.

- Britten, R. J. & McClure, F. T. The amino acid pool in *Escherichia coli*. *Bacteriol. Rev.* **26**, 292–335 (1962).
- Berrier, C., Coulombe, A., Szabo, I., Zoratti, M. & Ghazi, A. Gadolinium ion inhibits loss of metabolites induced by osmotic shock and large stretch-activated channels in bacteria. *Eur. J. Biochem.* **206**, 559–565 (1992).
- Martinac, B., Buechner, M., Delcour, A. H., Adler, J. & Kung, C. Pressure-sensitive ion channel in *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **84**, 2297–2301 (1987).
- Levina, N. *et al.* Protection of *Escherichia coli* cells against extreme turgor by activation of MscS and MscL mechanosensitive channels: identification of genes required for MscS activity. *EMBO J.* **18**, 1730–1737 (1999).
- Martinac, B., Delcour, A. H., Buechner, M., Adler, J. & Kung, C. *Advances in Comparative and Environmental Physiology* 3–18 (Springer, Heidelberg, 1992).
- Delcour, A. H., Martinac, B., Adler, J. & Kung, C. Modified reconstitution method used in patch-clamp studies of *Escherichia coli* ion channels. *Biophys. J.* **56**, 631–636 (1989).
- Sukharev, S. I., Blount, P., Martinac, B., Blattner, F. R. & Kung, C. A large-conductance mechanosensitive channel in *E. coli* encoded by *mscL* alone. *Nature* **368**, 265–268 (1994).
- Sukharev, S. I., Martinac, B., Blount, P. & Kung, C. Functional reconstitution as an assay for biochemical isolation of channel proteins: application to the molecular identification of a bacterial mechanosensitive channel. *Methods: A Companion to Methods in Enzymology* **6**, 51–59 (1994).
- Blount, P. *et al.* Membrane topology and multimeric structure of a mechanosensitive channel protein of *Escherichia coli*. *EMBO J.* **15**, 4798–4805 (1996).
- Chang, G., Spencer, R. H., Lee, A. T., Barclay, M. T. & Rees, D. C. Structure of the MscL homolog from *Mycobacterium tuberculosis*: a gated mechanosensitive ion channel. *Science* **282**, 2220–2226 (1998).
- Sukharev, S., Betanzos, M., Chiang, C. S. & Guy, H. R. The gating mechanism of the large mechanosensitive channel MscL. *Nature* **409**, 720–724 (2001).
- Perozo, E., Cortes, D. M., Sompornpisut, P., Kloda, A. & Martinac, B. Open channel structure of MscL and the gating mechanism of mechanosensitive channels. *Nature* **418**, 942–948 (2002).
- Perozo, E., Kloda, A., Cortes, D. M. & Martinac, B. Physical principles underlying the transduction of bilayer deformation forces during mechanosensitive channel gating. *Nature Struct. Biol.* **9**, 696–703 (2002).
- Bass, R. B., Strop, P., Barclay, M. & Rees, D. C. Crystal structure of *Escherichia coli* MscS, a voltage-modulated and mechanosensitive channel. *Science* **298**, 1582–1587 (2002).
- Sukharev, S. I., Blount, P., Martinac, B. & Kung, C. Mechanosensitive channels of *Escherichia coli*: the MscL gene, protein, and activities. *Annu. Rev. Physiol.* **59**, 633–657 (1997).
- Hamill, O. P. & Martinac, B. Molecular basis of mechanotransduction in living cells. *Physiol. Rev.* **81**, 685–740 (2001).
- Sukharev, S. & Corey, D. P. Mechanosensitive channels: multiplicity of families and gating paradigms. *Sci. STKE* doi:10.1126/stke.2332004eg7 (2004).
- Blount, P. Molecular mechanisms of mechanosensation: big lessons from small cells. *Neuron* **37**, 731–734 (2003).
- Sukharev, S. I., Sigurdson, W. J., Kung, C. & Sachs, F. Energetic and spatial parameters for gating of the bacterial large conductance mechanosensitive channel, MscL. *J. Gen. Physiol.* **113**, 525–540 (1999).
- Sukharev, S. I., Martinac, B., Arshavsky, V. Y. & Kung, C. Two types of mechanosensitive channels in the *Escherichia coli* cell envelope: solubilization and functional reconstitution. *Biophys. J.* **65**, 177–183 (1993).
- Ou, X. R., Blount, P., Hoffman, R. J. & Kung, C. One face of a transmembrane helix is crucial in mechanosensitive channel gating. *Proc. Natl Acad. Sci. USA* **95**, 11471–11475 (1998).
- Maurer, J. A. & Dougherty, D. A. Generation and evaluation of a large mutational library from the *Escherichia coli* mechanosensitive channel of large conductance, MscL: implications for channel gating and evolutionary design. *J. Biol. Chem.* **278**, 21076–21082 (2003).
- Cantor, R. S. Lateral pressures in cell membranes: a mechanism for modulation of protein function. *J. Phys. Chem.* **101**, 1723–1725 (1997).
- Cantor, R. S. The influence of membrane lateral pressures on simple geometric models of protein conformational equilibria. *Chem. Phys. Lipids* **101**, 45–56 (1999).
- Lindahl, E. & Edholm, O. Spatial and energetic-entropic decomposition of surface tension in lipid bilayers from molecular dynamics simulations. *J. Chem. Phys.* **113**, 3882–3893 (2000).
- Gullingsrud, J. & Schulten, K. Gating of MscL studied by steered molecular dynamics. *Biophys. J.* **85**, 2087–2099 (2003).
- Wiggins, P. & Phillips, R. Analytic models for mechanotransduction: gating a mechanosensitive channel. *Proc. Natl Acad. Sci. USA* **101**, 4071–4076 (2004).
- Sheetz, M. P. & Singer, S. J. Biological membranes as bilayer couples. A molecular mechanism of drug-erythrocyte interactions. *Proc. Natl Acad. Sci. USA* **71**, 4457–4461 (1974).
- Martinac, B., Adler, J. & Kung, C. Mechanosensitive ion channels of *E. coli* activated by amphipaths. *Nature* **348**, 261–263 (1990).
- Lundbaek, J. A., Maer, A. M. & Andersen, O. S. Lipid bilayer electrostatic energy, curvature stress, and assembly of gramicidin channels. *Biochemistry* **36**, 5695–5701 (1997).
- Cantor, R. S. The lateral pressure profile in membranes: a physical mechanism of general anesthesia. *Toxicol. Lett.* **100–101**, 451–458 (1998).
- Patel, A. J., Lazdunski, M. & Honore, E. Lipid and mechano-gated 2P domain K⁺ channels. *Curr. Opin. Cell Biol.* **13**, 422–428 (2001).
- Patel, A. J. *et al.* A mammalian two pore domain mechano-gated S-like K⁺ channel. *EMBO J.* **17**, 4283–4290 (1998).
- Maingret, F., Patel, A. J., Lesage, F., Lazdunski, M. & Honore, E. Lysophospholipids open the two-pore domain mechano-gated K⁺ channels TREK-1 and TRAAK. *J. Biol. Chem.* **275**, 10128–10133 (2000).
- Patel, A. J. & Honore, E. Anesthetic-sensitive 2P domain K⁺ channels. *Anesthesiology* **95**, 1013–1021 (2001).
- Chemin, J. *et al.* A phospholipid sensor controls mechanogating of the K⁺ channel TREK-1. *EMBO J.* **24**, 44–53 (2005).
- Chemin, J. *et al.* Lysophosphatidic acid-operated K⁺ channels. *J. Biol. Chem.* **280**, 4415–4421 (2005).
- Yang, X. C. & Sachs, F. Block of stretch-activated ion channels in *Xenopus* oocytes by gadolinium and calcium ions. *Science* **243**, 1068–1071 (1989).
- Hamill, O. P. & McBride, D. W. Jr The pharmacology of mechanogated membrane ion channels. *Pharmacol. Rev.* **48**, 231–252 (1996).
- Ermakov, Y. A., Averbakh, A. Z., Yusipovich, A. I. & Sukharev, S. Dipole potentials indicate restructuring of the membrane interface induced by gadolinium and beryllium ions. *Biophys. J.* **80**, 1851–1862 (2001).
- Suchyna, T. M. *et al.* Identification of a peptide toxin from *Grammostola spatulata* spider venom that blocks cation-selective stretch-activated channels. *J. Gen. Physiol.* **115**, 583–598 (2000); erratum *J. Gen. Physiol.* **117**, 371 (2001).
- Guharay, G. & Sachs, F. Stretch-activated single ion channel currents in tissue-cultured embryonic chick skeletal muscle. *J. Physiol. (Lond.)* **352**, 685–701 (1984).
- Suchyna, T. M. *et al.* Bilayer-dependent inhibition of mechanosensitive channels by neuroactive peptide enantiomers. *Nature* **430**, 235–240 (2004).
- Zhang, Y., Gao, F., Popov, V. L., Wen, J. W. & Hamill, O. P. Mechanically gated channel activity in cytoskeleton-deficient plasma membrane blebs and vesicles from *Xenopus* oocytes. *J. Physiol. (Lond.)* **523**, 117–130 (2000).
- Maroto, R. *et al.* The role of TRPC1 in forming the mechanosensitive cation channel in frog oocytes. *Nature Cell Biol.* **7**, 179–185 (2005).
- Morris, C. E. & Horn, R. Failure to elicit neuronal macroscopic mechanosensitive currents anticipated by single-channel studies. *Science* **251**, 1246–1249 (1991).
- Zhang, Y. & Hamill, O. P. On the discrepancy between whole-cell and membrane patch mechanosensitivity in *Xenopus* oocytes. *J. Physiol. (Lond.)* **523**, 101–115 (2000).

48. Ernststrom, G. G. & Chalfie, M. Genetics of sensory mechanotransduction. *Annu. Rev. Genet.* **36**, 411–453 (2002).
49. Goodman, M. B. & Schwarz, E. M. Transducing touch in *Caenorhabditis elegans*. *Annu. Rev. Physiol.* **65**, 429–452 (2003).
50. Bianchi, L. *et al.* The neurotoxic MEC-4(d) DEG/ENaC sodium channel conducts calcium: implications for neurosis initiation. *Nature Neurosci.* **7**, 1337–1344 (2004).
51. Goodman, M. B. *et al.* MEC-2 regulates *C. elegans* DEG/ENaC channels needed for mechanosensation. *Nature* **415**, 1039–1042 (2002).
52. Chelur, D. S. *et al.* The mechanosensory protein MEC-6 is a subunit of the *C. elegans* touch-cell degenerin channel. *Nature* **420**, 669–673 (2002).
53. Ertage, L., Gu, G., Hartwig, E. & Chalfie, M. Extracellular proteins organize the mechanosensory channel complex in *C. elegans* touch receptor neurons. *Neuron* **44**, 795–807 (2004).
54. Zhang, S. *et al.* MEC-2 is recruited to the putative mechanosensory complex in *C. elegans* touch receptor neurons through its stomatic-like domain. *Curr. Biol.* **14**, 1888–1896 (2004).
55. O'Hagan, R., Chalfie, M. & Goodman, M. B. The MEC-4 DEG/ENaC channel of *Caenorhabditis elegans* touch receptor neurons transduces mechanical signals. *Nature Neurosci.* **8**, 43–50 (2005).
56. Tsunoda, S. *et al.* A multivalent PDZ-domain protein assembles signalling complexes in a G-protein-coupled cascade. *Nature* **388**, 243–249 (1997).
57. Montell, C. TRP trapped in fly signalling web. *Curr. Opin. Neurobiol.* **8**, 389–397 (1998).
58. Montell, C., Birnbaumer, L. & Flockerzi, V. The TRP channels, a remarkably functional family. *Cell* **108**, 595–598 (2002).
59. Corey, D. P. New TRP channels in hearing and mechanosensation. *Neuron* **39**, 585–588 (2003).
60. Minke, B., Wu, C.-F. & Pak, W. L. Induction of photoreceptor voltage noise in the dark in *Drosophila* mutant. *Nature* **258**, 84–87 (1975).
61. Colbert, H. A., Smith, T. L. & Bargmann, C. I. OSM-9, a novel protein with structural similarity to channels, is required for olfaction, mechanosensation, and olfactory adaptation in *Caenorhabditis elegans*. *J. Neurosci.* **17**, 8259–8269 (1997).
62. Barr, M. M. & Sternberg, P. W. A polycystic kidney-disease gene homologue required for male mating behaviour in *C. elegans*. *Nature* **401**, 386–389 (1999).
63. Walker, R. G., Willingham, A. T. & Zuker, C. S. A *Drosophila* mechanosensory transduction channel. *Science* **287**, 2229–2234 (2000).
64. Tracey, W. D. Jr, Wilson, R. I., Laurent, G. & Benzer, S. *painless*, a *Drosophila* gene essential for nociception. *Cell* **113**, 261–273 (2003).
65. Di Palma, F. *et al.* Mutations in Mcoln3 associated with deafness and pigmentation defects in varitint-waddler (Va) mice. *Proc. Natl Acad. Sci. USA* **99**, 14994–14999 (2002).
66. Mochizuki, T. *et al.* PKD2, a gene for polycystic kidney disease that encodes an integral membrane protein. *Science* **272**, 1339–1342 (1996).
67. Suzuki, M., Mizuno, A., Kodaira, K. & Imai, M. Impaired pressure sensation in mice lacking TRPV4. *J. Biol. Chem.* **278**, 22664–22668 (2003).
68. Gao, X., Wu, L. & O'Neil, R. G. Temperature-modulated diversity of TRPV4 channel gating: activation by physical stresses and phorbol ester derivatives through protein kinase C-dependent and -independent pathways. *J. Biol. Chem.* **278**, 27129–27137 (2003).
69. Liedtke, W. *et al.* Vanilloid receptor-related osmotically activated channel (VR-OAC), a candidate vertebrate osmoreceptor. *Cell* **103**, 525–535 (2000).
70. Strotmann, R., Harteneck, C., Nunnenmacher, K., Schultz, G. & Plant, T. D. OTRPC4, a nonselective cation channel that confers sensitivity to extracellular osmolarity. *Nature Cell Biol.* **2**, 695–702 (2000).
71. Tobin, D. *et al.* Combinatorial expression of TRPV channel proteins defines their sensory functions and subcellular localization in *C. elegans* neurons. *Neuron* **35**, 307–318 (2002).
72. Kahn-Kirby, A. *et al.* Specific polyunsaturated fatty acids drive TRPV-dependent sensory signalling *in vivo*. *Cell* **119**, 889–900 (2004).
73. Liedtke, W., Tobin, D. M., Bargmann, C. I. & Friedman, J. M. Mammalian TRPV4 (VR-OAC) directs behavioural responses to osmotic and mechanical stimuli in *Caenorhabditis elegans*. *Proc. Natl Acad. Sci. USA* **100** (suppl. 2), 14531–14536 (2003).
74. Watanabe, H. *et al.* Anandamide and arachidonic acid use epoxyeicosatrienoic acids to activate TRPV4 channels. *Nature* **424**, 434–438 (2003).
75. Caterina, M. J. *et al.* The capsaicin receptor: a heat-activated ion channel in the pain pathway. *Nature* **389**, 816–824 (1997).
76. Birder, L. A. *et al.* Altered urinary bladder function in mice lacking the vanilloid receptor TRPV1. *Nature Neurosci.* **5**, 856–860 (2002).
77. Vriens, J. *et al.* Cell swelling, heat, and chemical agonists use distinct pathways for the activation of the cation channel TRPV4. *Proc. Natl Acad. Sci. USA* **101**, 396–401 (2004).
78. Kim, J. *et al.* A TRPV family ion channel required for hearing in *Drosophila*. *Nature* **424**, 81–84 (2003).
79. Gong, S. *et al.* Two interdependent TRPV channel subunits, inactive and Nanchung, mediate hearing in *Drosophila*. *J. Neurosci.* **24**, 9059–9066 (2004).
80. Sidi, S., Friedrich, R. W. & Nicolson, T. NompC TRP channel required for vertebrate sensory hair cell mechanotransduction. *Science* **301**, 96–99 (2003).
81. Siemens, J. *et al.* Cadherin 23 is a component of the tip link in hair-cell stereocilia. *Nature* **428**, 950–955 (2004).
82. Sollner, C. *et al.* Mutations in cadherin 23 affect tip links in zebrafish sensory hair cells. *Nature* **428**, 955–959 (2004).
83. Corey, D. P. *et al.* TRPA1 is a candidate for the mechanosensitive transduction channel of vertebrate hair cells. *Nature* **432**, 723–730 (2004).
84. Howard, J. & Bechstedt, S. Hypothesis: a helix of ankyrin repeats of the NOMP-TRP ion channel is the gating spring of mechanoreceptors. *Curr. Biol.* **14**, R224–R226 (2004).
85. Jiang, Y. *et al.* The open pore conformation of potassium channels. *Nature* **417**, 523–526 (2002).
86. Erler, I., Hirnet, D., Wissenbach, U., Flockerzi, V. & Niemeyer, B. A. Ca²⁺-selective transient receptor potential V channel architecture and function require a specific ankyrin repeat. *J. Biol. Chem.* **279**, 34456–34463 (2004).
87. Hirono, M., Denis, C. S., Richardson, G. P. & Gillespie, P. G. Hair cells require phosphatidylinositol 4,5-bisphosphate for mechanical transduction and adaptation. *Neuron* **44**, 309–320 (2004).
88. Story, G. M. *et al.* ANKTM1, a TRP-like channel expressed in nociceptive neurons, is activated by cold temperatures. *Cell* **112**, 819–829 (2003).
89. Jordt, S. E. *et al.* Mustard oils and cannabinoids excite sensory nerve fibres through the TRP channel ANKTM1. *Nature* **427**, 260–265 (2004).
90. Bandell, M. *et al.* Noxious cold ion channel TRPA1 is activated by pungent compounds and bradykinin. *Neuron* **41**, 849–857 (2004).
91. Palmer, C. P. *et al.* A TRP homolog in *Saccharomyces cerevisiae* forms an intracellular Ca²⁺-permeable channel in the yeast vacuolar membrane. *Proc. Natl Acad. Sci. USA* **98**, 7801–7805 (2001).
92. Zhou, X. L. *et al.* The transient receptor potential channel on the yeast vacuole is mechanosensitive. *Proc. Natl Acad. Sci. USA* **100**, 7105–7110 (2003).
93. Denis, V. & Cyert, M. S. Internal Ca²⁺ release in yeast is triggered by hypertonic shock and mediated by a TRP channel homologue. *J. Cell Biol.* **156**, 29–34 (2002).
94. Zhou, X.-L., Loukin, S. H., Coria, R., Kung, C. & Saimi, Y. Heterologously expressed fungal transient receptor potential channels retain mechanosensitivity *in vitro* and osmotic response *in vivo*. *Eur. Biophys. J.* **34**, 413–422 (2005).
95. Kung, C., Saimi, Y. & Martinac, B. *Current Topics in Membranes and Transport* 145–153 (Academic, New York, 1990).
96. Apostle, H. G. *Aristotle's On The Soul (De Anima) (Translation)* 42–43 (Peripatetic, Crinnell, Iowa, 1981).
97. Clapham, D. E. TRP channels as cellular sensors. *Nature* **426**, 517–524 (2003).
98. Voets, T. *et al.* The principle of temperature-dependent gating in cold- and heat-sensitive TRP channels. *Nature* **430**, 748–754 (2004).
99. Corey, D. P. & Sotomayor, M. Hearing: tightrope act. *Nature* **428**, 901–903 (2004).
100. Kung, C. in *Evolution of the First Nervous Systems* (ed. Anderson, P. A. V.) 203–214 (Plenum, New York, 1990).

Acknowledgements I thank A. Anishkin, M. Chalfie, R. Fettiplace, W. J. Haynes, S. Loukin, B. Martinac, Y. Saimi, A. O. W. Stretton and X.-L. Zhou for discussions and criticisms. My laboratory is supported by the Vilas Trust of the University of Wisconsin and by the NIH.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to C.K. (ckung@wisc.edu).

Terrestrial nitrogen and noble gases in lunar soils

M. Ozima¹, K. Seki², N. Terada^{2†}, Y. N. Miura³, F. A. Podosek⁴ & H. Shinagawa^{2†}

The nitrogen in lunar soils is correlated to the surface and therefore clearly implanted from outside. The straightforward interpretation is that the nitrogen is implanted by the solar wind, but this explanation has difficulties accounting for both the abundance of nitrogen and a variation of the order of 30 per cent in the $^{15}\text{N}/^{14}\text{N}$ ratio. Here we propose that most of the nitrogen and some of the other volatile elements in lunar soils may actually have come from the Earth's atmosphere rather than the solar wind. We infer that this hypothesis is quantitatively reasonable if the escape of atmospheric gases, and implantation into lunar soil grains, occurred at a time when the Earth had essentially no geomagnetic field. Thus, evidence preserved in lunar soils might be useful in constraining when the geomagnetic field first appeared. This hypothesis could be tested by examination of lunar farside soils, which should lack the terrestrial component.

Since the Apollo missions, it has been recognized that the Moon is very strongly depleted in volatile elements, including N, H, C and the noble gases^{1,2}. The inventory of these elements in lunar materials is not intrinsically lunar but rather reflects an extralunar origin. The extralunar source is generally understood to be the Sun, via direct implantation of solar wind (SW) ions in the surface of the Moon. This interpretation suffices well for the other volatile elements, but encounters difficulty with N—that is, there is too much N compared to the canonical solar abundance (of, for example, Ar), and the isotopic composition ($^{15}\text{N}/^{14}\text{N}$) of surface-correlated and presumably extralunar N is strongly variable, by as much as 30%. The overabundance of N has been attributed to underabundance of noble gases such as Ar (refs 1, 2), and the isotopic variation to the variation of N composition in the source region of the SW.

However, quantitative evaluations³ preclude any suggested mechanism by which the isotopic composition of N could vary so much in either the photosphere or the SW. This has led in turn to hypotheses in which some or most surface-correlated lunar N is extrasolar as well as extralunar: that is, it reflects a source other than the SW. Potential sources that have been suggested include interstellar gas, intrinsic lunar N degassed from the interior, the terrestrial atmosphere and infall of cometary or asteroidal debris. The extrasolar interpretation has been strengthened by recent experimental results^{4,5}. Overall, no specific model has gained a clear ascendancy.

Here we propose that sources of the extrasolar N and light noble gases in lunar soils can be attributed to ion flows from the atmosphere of a non-magnetic Earth. Mechanisms responsible for the atmospheric escape from a planet depend on its magnetic field^{6,7}. An atmospheric source of extrasolar N in lunar soils was suggested in a framework of ion implantation in Earth's magnetotail^{3,8}. However, recent observations of the O^+ escape flux⁷ and the N^+/O^+ ratio⁹ indicate that the N^+ flux at the Moon is less than $10^3 \text{ cm}^{-2} \text{ s}^{-1}$, which is insufficient to account for the extrasolar N flux in lunar soils. However, if the geomagnetic field (GMF) were absent, the SW would directly interact with the upper atmosphere, causing a much larger ion escape flux.

The origin of the GMF is still not well understood; in particular,

the fundamental issue of when it first appeared remains enigmatic. The oldest palaeomagnetic information available is based on palaeointensity measurements on the Komati formation (age 3.5 Gyr)¹⁰, which showed a much smaller virtual geomagnetic dipole moment than the present value. (But also note that this conclusion is not unchallenged¹¹.)

If the origin of the GMF were concomitant with the formation of a liquid core, the age of the appearance of the GMF would probably be essentially the same as that of the Earth, earlier than the formation of any significant fraction of the present regolith. For such an early appearance of the GMF, we could thus not expect preservation of any record in lunar soils. On the other hand, if the appearance of the GMF were significantly delayed, it is possible that there was a significant window when the Earth had no GMF and the Moon had a stable regolith surface that could preserve implanted ions to the present day. In that case, the onset of the geomagnetic dynamo might be recorded in lunar soils.

Ion escape from a non-magnetic Earth

In the presence of the GMF, the SW approaching the Earth is stopped at the magnetopause (at a distance of ~ 10 Earth radii, $10R_E$), where the GMF magnetic pressure balances the SW dynamic pressure. If the GMF did not exist or were much weaker, the SW would approach much closer to the Earth, until its dynamic pressure was balanced by ion pressure at the ionopause, about 500 km above the surface. The SW must thus have interacted directly with the ionized (by solar radiation) upper atmosphere. Atmospheric ions around the ionopause could then have been picked up by the incoming SW and carried away from the Earth⁶. This SW-induced ion escape mechanism from a non-magnetic planet has been observed for Venus¹², which has no permanent magnetic dipole. Thus, if the GMF were absent or extremely weak, we would expect substantial ion escape from the Earth's ionosphere. Some fraction of the escaping ions would have hit the Moon.

Pioneer Venus Orbiter (PVO) observed an escape flux of $\sim 5 \times 10^7 \text{ O}^+ \text{ ions cm}^{-2} \text{ s}^{-1}$ from the venusian atmosphere¹², and $\text{N}^+/\text{O}^+ \approx 0.02$ at the ionopause¹³. The escape of low-energy

¹Graduate School of Earth and Planetary Science, University of Tokyo, Tokyo 113-0033, Japan. ²Solar-Terrestrial Environment Laboratory, Nagoya University, Honohara 3-13, Toyokawa, Aichi 442-8507, Japan. ³Earthquake Research Institute, University of Tokyo, Tokyo 113-0032, Japan. ⁴Institute of Isotope Geology and Mineral Resources, ETH-Zentrum, 8092 Zürich, Switzerland, and Department of Earth Planetary Sciences, Washington University, St Louis, Missouri 63130-4893, USA. [†]Present address: National Institute of Information and Communications Technology, Tokyo 184-8795, Japan.

ions through the iontail suggests a magnetohydrodynamic type of ion outflow, independent of the ion mass, and we infer an N^+ escape flux of $10^6 \text{ ions cm}^{-2} \text{ s}^{-1}$ from the venusian atmosphere. It is commonly assumed that the major constituent of the primitive Earth's atmosphere (before the growth of biogenic O) was CO_2 (ref. 14). Therefore, if the ancient Earth did not have a GMF, we expect an escape flux of N from the ancient Earth similar to that from Venus.

To examine ion escape fluxes from a non-magnetic Earth quantitatively, we have made numerical calculations to estimate ion escape fluxes of N^+ and the light noble gases He^+ , Ne^+ and Ar^+ from a presumed non-magnetic Earth. These calculations were made for the present Earth's upper atmosphere from the MSIS00 model¹⁵, with the ion production rate given by ref. 16. We assume that all ions produced above the ionopause are picked up by the SW and escape. Fluxes were calculated by integrating the total ion production rate over the dayside Earth surface area above the ionopause, normalized by assuming a circular escape area of radius $2R_E$ (see below). Figure 1 shows thus-calculated ion escape rates against ionopause altitude (and corresponding SW dynamic pressure, P_{SW}). If we assume the ionopause altitude of 500 km, which is inferred for a non-magnetic Earth under typical present SW conditions, the calculated ion escape rates for N^+ will be $\sim 10^6 \text{ ions cm}^{-2} \text{ s}^{-1}$ (Fig. 1).

This calculated N^+ escape rate is comparable with the N^+ escape flux ($\sim 10^6 \text{ ions cm}^{-2} \text{ s}^{-1}$) inferred from the PVO. Figure 1 also shows that if the SW dynamic pressure changes from 3.5 to 22 nPa, the ionopause altitude decreases from $\sim 550 \text{ km}$ down to $\sim 250 \text{ km}$, and accordingly the ion escape rates, especially of heavy ions such as Ne^+ and Ar^+ , undergo drastic increase. For example, the escape rate of N^+ increases from 4×10^5 to 2×10^8 , Ne^+ from 60 to 6×10^4 , and $^{36}\text{Ar}^+$ from 0.004 to $6 \times 10^2 \text{ ions cm}^{-2} \text{ s}^{-1}$, respectively. The ion escape flux depends on the scale height of each ion species and increases nearly exponentially with the ionopause altitude, but not with P_{SW} . As the ionopause altitude becomes lower, higher P_{SW} is needed to further decrease the ionopause altitude, as labelled at the right-hand side in the figure, which is based on the altitude profile of the ionospheric pressure.

If the activity of a young Sun were higher than at present and continuous flare events enhanced SW dynamic pressure, the escape rate would have been much higher than the above estimation based on the present Earth's condition. It is also worth noting that the primitive Earth's atmosphere must have been considerably depleted in oxygen relative to the present Earth's atmosphere, which should have brought the ionopause closer to the Earth to increase the ion escape rate. We show below (Table 1 and discussions) that the origin

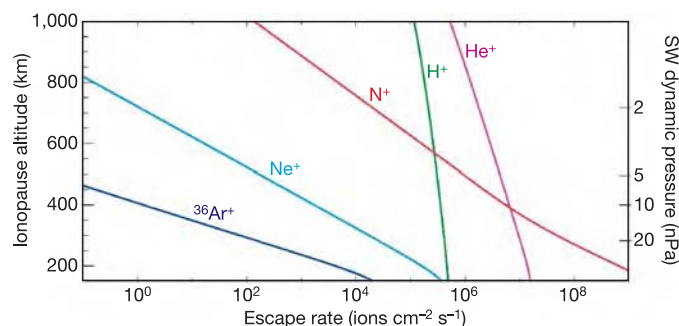


Figure 1 | Ion escape rates versus ionopause altitude for the present Earth's atmosphere without a global magnetic field. Escape rates of $^{36}\text{Ar}^+$ (blue), Ne^+ (light blue), N^+ (red), He^+ (magenta) and H^+ (green) ions were estimated using the NRLMSISE00 model of the present Earth's upper neutral atmosphere¹⁵ and ionization rates¹⁶ for normal extreme-ultraviolet flux ($F_{10.7} = 150 \times 10^{-22} \text{ J s}^{-1} \text{ m}^{-2} \text{ Hz}^{-1}$). The ordinate represents both the ionopause altitude (left) and the corresponding solar wind (SW) dynamic pressure (right) derived from the IRI model³².

of non-solar components of N and light noble gases in lunar soils can reasonably be attributed to the Earth's ionosphere.

Assuming an N^+ escape flux of $10^6 \text{ ions cm}^{-2} \text{ s}^{-1}$ from the Earth, we next estimate the fraction of the escaping ions that will hit the lunar surface. In analogy to the PVO observation, we assumed a circular cross-section of $2R_E$ for the escaping area of terrestrial ions (hereafter Earth wind or EW). In evaluating the probability of the Moon's passage through the EW area, we also considered the variation of the Earth–Moon distance with time; the distance between the Earth and the Moon has been increasing owing to tidal dissipation since the formation of the Earth–Moon system^{17,18}. Figure 2a shows a recent estimation of the time variation of the E–M distance¹⁸. For an Earth–Moon distance of $40R_E$, corresponding to about 4 Gyr ago (Fig. 2a), we infer from Fig. 2b that on average about 0.3% of the EW flux would hit the lunar surface, that is, about $3 \times 10^3 \text{ N}^+ \text{ cm}^{-2} \text{ s}^{-1}$. The N^+ ions will then be implanted in lunar soils with the same velocity as that of the shocked SW of $\sim 440 \text{ km s}^{-1}$ ($\sim 1 \text{ keV}$ per a.m.u.). As will be shown later (Table 1), the above estimated average N flux (about $3 \times 10^3 \text{ atoms cm}^{-2} \text{ s}^{-1}$) that could be transported to the Moon from a non-magnetic Earth is close to the non-solar N flux observed at the lunar surface ($> 2 \times 10^3 \text{ atoms cm}^{-2} \text{ s}^{-1}$).

Non-solar components in lunar soils

Nitrogen. Hashizume *et al.*¹⁹ argued that the isotopic variation in surface-correlated lunar N can best be attributed to the mixing of two distinct components, solar (SW) and a putative planetary

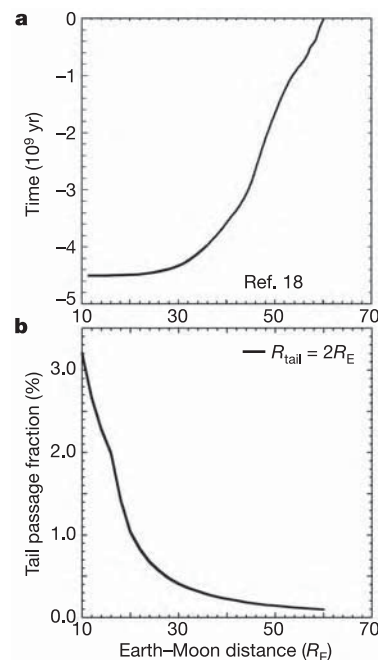


Figure 2 | Estimation of the fraction of the Earth wind (EW) that reaches the Moon. **a**, Earth–Moon distance (in units of Earth radius, R_E) versus age (redrawn from ref. 18); the present day is at time zero. **b**, The time that the Moon spends within the EW tail as a function of the Earth–Moon distance (abscissa). In the ecliptic (x – y) plane, the probability P_{x-y} of the EW tail shadowing the Moon is $\sin^{-1}(S/R)/2\pi$, where R is the distance between the Earth and the Moon, and S is the radius of the EW tail. In estimating the probability P_z of the tail shadowing the Moon in the z -direction, we assume that the Moon's orbital plane is inclined to the ecliptic by $i = 6.7^\circ$ (the present value). Probability P_z then becomes $P(|R\sin(i)\cos(x:0 \leq x < 2\pi)| < S)$. For the EW flux, we assume a circular cylinder of $2R_E$ radius, as suggested by PVO observation. The total probability $P = P_{x-y} \times P_z$ is shown as a function of R (see **a**). The kink in the curve at $\sim 17R_E$ corresponds to the distance where the lunar orbit starts to exit from the cylindrical EW tail.

Table 1 | SW and EW fluxes in lunar soils

	Solar abundance at Moon surface	SW flux ($\text{cm}^{-2} \text{s}^{-1}$)	Non-solar comp. (fraction)	Non-solar flux ($\text{cm}^{-2} \text{s}^{-1}$)	EW flux (estimated) ($\text{cm}^{-2} \text{s}^{-1}$)	EW flux (calculated) ($\text{cm}^{-2} \text{s}^{-1}$)
^4He	1.0	6.3×10^6	~ 0.3	2×10^6	6.7×10^8	2×10^7
^{20}Ne	1.54×10^{-3}	9.7×10^3	~ 0.3	3×10^3	1.0×10^6	10^5
^{36}Ar	3.27×10^{-5}	206	~ 0.05	10	3.3×10^3	10^3
$^{14}\text{N}^*$		4×10^3	> 0.5	2×10^3	$> 7 \times 10^5$	10^8

SW ^4He flux at Moon surface (Al foil observation) = $6.3 \times 10^6 \text{ cm}^{-2} \text{s}^{-1}$ (ref. 21). Column 1, SW relative elemental (X) abundance at Moon surface = $[X]/[^4\text{He}]$ (ref. 33). Column 2, SW flux (element X) at Moon surface = $[O] \times [1]$. Column 3, Fraction of non-solar component (X) in lunar soil estimated from a mixing diagram (Figs 3 and 4). Column 4, Non-solar X flux in lunar soil = $[2] \times [3]$. Column 5, Earth Wind (EW) flux = $[4]/0.003$, if non-solar X were totally attributed to EW (see text). Column 6, EW flux calculated for potential palaeoatmosphere. The height of ionopause is assumed to be 250 km.

*Data for N are from refs 5, 19.

component that they assigned to interplanetary dust particles (IDPs) (Fig. 3). However, other than isotopically this putative IDP component is not well constrained. It may be noted that the N and H isotopic compositions of the assumed planetary component are close to the terrestrial atmospheric compositions, so that the mixing trend suggested in Fig. 3 might be equally well attributed to mixing between SW and terrestrial components.

Light noble gases. If a substantial amount of N can escape from a non-magnetic Earth, we would expect that other volatile elements also escape from the ionopause. Recent experimental observations on light noble gases (He, Ne, Ar) in lunar soils suggest this possibility: Heber *et al.*²⁰ analysed ilmenite separates (grain ensembles, but not single grains) from Apollo lunar soils (71501, 12001, 74241) and regolith breccias (79135, 79035, 14301) for light noble gases. They found that the isotopic compositions of He, Ne and Ar show significant variation, but also that they are well correlated with each other. As they further pointed out, the correlation clearly indicates that variation in He isotopic ratio cannot be attributed to nuclear process in the Sun, as is often suggested in the literature, since no nuclear process in the Sun can quantitatively affect Ne isotopic ratios. Heber *et al.*²⁰ therefore suggested that the good correlation between isotopic ratios was due to some experimental artefact. However, we suggest that the observed correlation, not only between

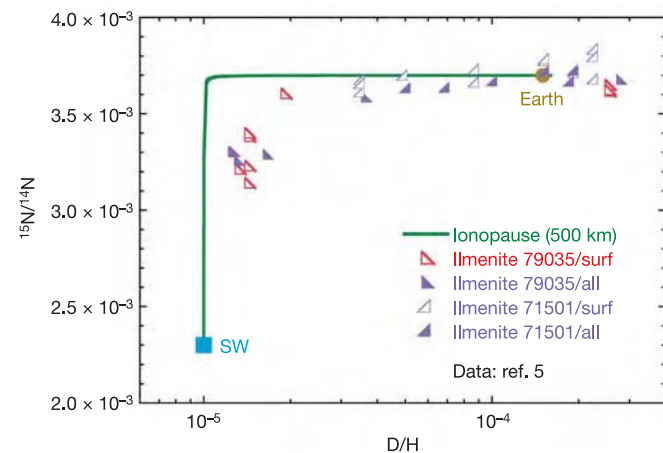


Figure 3 | D/H- $^{15}\text{N}/^{14}\text{N}$ mixing diagram. Experimental data (red and purple triangles) for individual ilmenite grains separated from Apollo 17 breccias (79035 and 71501) obtained in ref. 5 are plotted in a $^{15}\text{N}/^{14}\text{N}$ -D/H diagram ('surf' indicates grain surface; 'all' indicates whole grain). A mixing curve between a SW component (blue square) and a terrestrial component (olive circle) was constructed with three independently estimated parameters (two end member isotopic compositions and the ratio of elemental ratios of the two end members, that is, $r = (^{14}\text{N}/\text{H})_{\text{E}}/(^{14}\text{N}/\text{H})_{\text{SW}}$). We followed refs 5 and 19 for the SW composition. For terrestrial components, we used the present atmospheric ratios, $\text{D}/\text{H} = 1.5 \times 10^{-4}$ and $^{15}\text{N}/^{14}\text{N} = 3.7 \times 10^{-3}$. The mixing curve is for ionopause height 500 km (but is indistinguishable from the case for 300 km). A ratio of $(^{14}\text{N}/\text{H})_{\text{E}}$ (at 500 km) is calculated from Fig. 1.

$^3\text{He}/^4\text{He}$ and $^{20}\text{Ne}/^{22}\text{Ne}$, but also between $^3\text{He}/^4\text{He}$ and $^{40}\text{Ar}/^{36}\text{Ar}$ (the latter correlation was also reported by Geiss²¹ for bulk regolith samples) can be well explained on the basis of mixing between SW components and terrestrial atmospheric components.

In Fig. 4a, average $^3\text{He}/^4\text{He}$ and $^{20}\text{Ne}/^{22}\text{Ne}$ ratios obtained by Heber *et al.*²⁰ on six lunar ilmenite separates are reproduced. The data points can be adequately represented by a mixing curve between SW and terrestrial components, where the mixing curve is constructed with three independently estimated mixing parameters and for ionopause altitudes of 500 km, 350 km (shown only for the He-Ne

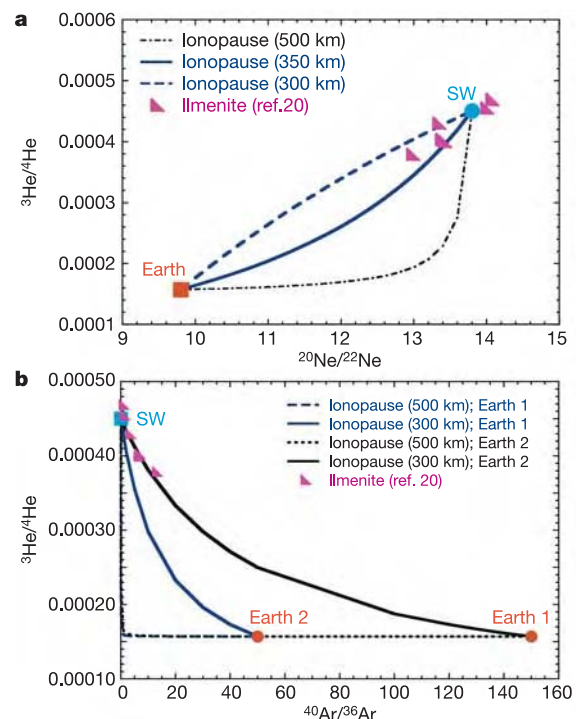


Figure 4 | Mixing diagrams. Shown are mixing diagrams between $^3\text{He}/^4\text{He}$ and $^{20}\text{Ne}/^{22}\text{Ne}$ (a), and between $^3\text{He}/^4\text{He}$ and $^{40}\text{Ar}/^{36}\text{Ar}$ (b). Data (filled triangles) are obtained for ilmenite grain ensemble separated from lunar soils (12001, 71501, 74241) and from lunar breccias (14301, 79135, 79035) (ref. 20). **a**, For the SW component (blue circle), we choose the generally assumed values³³ of $^3\text{He}/^4\text{He} = 4.5 \times 10^{-4}$, $^{20}\text{Ne}/^{22}\text{Ne} = 13.8$ and $^{22}\text{Ne}/^4\text{He} = 1.12 \times 10^{-4}$. For the Earth components (orange square), we use the present atmospheric ratio ($^{20}\text{Ne}/^{22}\text{Ne} = 9.8$) and the primordial solar ratio ($^3\text{He}/^4\text{He} = 1.57 \times 10^{-4}$; see text). Mixing curves were constructed for ionopause altitudes 500 km, 350 km and 300 km with respective $(^{22}\text{Ne}/^4\text{He})_{\text{E}}$ ratios estimated from Fig. 1. **b**, For the SW component (blue square), we choose the generally assumed values³² of $^3\text{He}/^4\text{He} = 4.5 \times 10^{-4}$, $^{40}\text{Ar}/^{36}\text{Ar} = 10^{-4}$ and $^{36}\text{Ar}/^4\text{He} = 3.27 \times 10^{-5}$. For the Earth components (orange circles), we assumed $^3\text{He}/^4\text{He} = 1.57 \times 10^{-4}$ for the primordial terrestrial ratio (see text). We constructed mixing curves for $^{40}\text{Ar}/^{36}\text{Ar} = 50$ (Earth 2) and $^{40}\text{Ar}/^{36}\text{Ar} = 150$ (Earth 1), and for ionopause altitudes of 500 km and 300 km, with respective $(^{36}\text{Ar}/^4\text{He})_{\text{E}}$ ratios calculated from Fig. 1.

case) and 300 km (see Fig. 4 legend). Among the three mixing parameters, the choice of $^3\text{He}/^4\text{He}$ for the ancient atmosphere deserves special attention, since there are two prominent He components in the early Solar System, namely pre-D-burning He ($^3\text{He}/^4\text{He} = 1.4 \times 10^{-4}$) and post-D-burning He ($^3\text{He}/^4\text{He} = 4.5 \times 10^{-4}$). The former is observed in Jupiter's atmosphere²² and also in the so-called Q component²³ in primitive meteorites. The latter represents He in the present SW (it differs from the primordial value because of creation of ^3He by nuclear 'burning' of D in the pre-main-sequence proto-Sun) and is also observed in some meteorites²⁴. From a visual inspection of the mixing diagrams, we note that if the distribution were due to the mixing of the SW and the Earth components, the pre-D-burning He can fairly well be assigned as the Earth end member.

In constructing a mixing curve between SW and terrestrial components for He–Ar, we therefore tentatively used the solar isotopic ratios for SW ($^3\text{He}/^4\text{He} = 4.5 \times 10^{-4}$ and $^{40}\text{Ar}/^{36}\text{Ar} = 10^{-4}$) and the pre-D-burning He isotopic ratio for the primitive Earth. However, the choice of appropriate $^{40}\text{Ar}/^{36}\text{Ar}$ for the ancient Earth atmosphere is difficult because of the extremely rapid evolution in this ratio in the terrestrial atmosphere owing to the addition of radiogenic ^{40}Ar (ref. 25). (For example, the present atmospheric $^{40}\text{Ar}/^{36}\text{Ar}$ (295.5) is more than six orders of magnitude larger than the primordial cosmic ratio (10^{-4})). We examined mixing curves for two cases, that is, $(^{40}\text{Ar}/^{36}\text{Ar})_{\text{E}} = 50$ and 150; here subscript E indicates Earth.

Both He–Ne and He–Ar mixing curves (Fig. 4a, b) appear to show that the pre-D-burning He may be assigned to the end member corresponding to the Earth. A straightforward implication would then be that the Earth inherited the pre-D-burning He when it formed. However, this interpretation requires that the contribution of radiogenic ^4He had not been important in the first 0.5 Gyr or so of the period in which ilmenite grains were likely to be implanted by the EW (see below for the estimation of the SW exposure age). If $\text{U}(\text{Th})/^3\text{He}$ were large in the Earth, the primordial post-D-burning He might have evolved to the isotopic ratio assumed for the Earth in the mixing diagrams. From the mixing diagrams alone, we cannot rule out either possibility, that is, pre-D-burning or the post-D-burning He as possible primordial He in the Earth.

Implanted EW components

Table 1 shows non-solar N and light noble gas fluxes estimated in lunar soils (details of estimation procedures are also given). From a comparison with curves for various ions (Fig. 1) for an ionopause altitude of 250 km, which is reasonable to assume for the primitive abiogenic atmosphere, we infer that EW can account for non-solar N and Ar (assuming $^{40}\text{Ar}/^{36}\text{Ar} = 150$) and barely for Ne (about 10%). If the SW flux in ancient times was more intense, which is likely, most of the Ne may also be accounted for by EW.

The SW sweeps ions away above the ionopause regardless of their mass, and would not cause isotopic fractionation of its constituents. The pick-up EW ions are quickly accelerated to the speed of the SW, and some are then implanted in lunar soils with the same velocity as that of the SW. Therefore, it would be difficult to discriminate between implanted EW components from SW components on the basis of the depth profile of implantation in lunar soils. Although ions would not undergo isotopic fractionation once picked up by the SW, isotopic compositions of pick-up ions may reflect fractionation that took place below the ionopause.

Lunar soil, a tracer of GMF evolution?

In order to use lunar soil as a tracer of ancient GMF variation, we must know the time when terrestrial atmospheric components were implanted in lunar soils. If we find that the majority of lunar soils older than a certain age were systematically endowed with a disproportionately large fraction of plausible terrestrial components, we may conclude that GMF did not exist before this age.

Several authors have suggested that the $^{40}\text{Ar}/^{36}\text{Ar}$ ratio may be used as an antiquity parameter for the surface exposure of lunar soil²⁶. However, Ozima *et al.*²⁷ questioned the validity of this method, since the assumed process (degassing from lunar interior and re-implanting in soil) requires an unrealistically large Ar degassing rate from the lunar interior. Relevant information regarding surface exposure age of lunar soils may be obtained by the use of cosmic ray and surface neutron irradiation effects, but this only tells the time during which samples were within a few metres of the surface, not the exposure time of ilmenite grains directly exposed to the SW. Several samples from Apollo 14 and from Apollo 17 have been studied for the relevant cosmic ray and surface neutron irradiation data^{26,28}. Although it is difficult to relate the irradiation age of lunar soil to the surface implantation age of an individual ilmenite grain in the soil, the conclusion by Bernatowicz *et al.*²⁸ may be worth noting: they stated "the extreme case that the irradiation age coincided with the formation age (of minerals in soils) was most easily in accord with the data at hand". Hence, it may be possible that the substantial fractions of ilmenite grains used in this study had a surface implantation age close to 3.8–3.9 Gyr ago, which is generally assigned to the formation age^{29,30} of major impact basins from where the Apollo samples were collected. It is then tempting to speculate that the GMF was null or very weak before about 3.9 Gyr ago.

Test of the hypothesis

Observed data so far available in the literature appear to be consistent with our hypothesis on EW implantation in lunar soils. We suggest that a further test of this hypothesis could be effected by comparing N and light noble gas data for lunar soil/breccia samples from the nearside of the Moon with those from the farside. The Moon–Earth dynamic system, once formed, is thought to have become very promptly (within a few tens of Myr) spin-locked owing to tidal interactions³¹. Thus, only the present nearside of the Moon has been facing the Earth throughout most of the history of the system, and so the farside of the Moon should be essentially free from EW material.

Received 16 November 2004; accepted 9 June 2005.

- Kerridge, J. F. Solar nitrogen: Evidence for a secular increase in the ratio of nitrogen-15 to nitrogen-14. *Science* **188**, 162–164 (1975).
- Kerridge, J. F. Corpuscular radiation: Evidence from nitrogen isotopes in the lunar regolith. *Rev. Geophys.* **31**, 423–437 (1993).
- Geiss, J. & Bochsler, P. Nitrogen isotopes in the solar system. *Geochim. Cosmochim. Acta* **46**, 529–548 (1982).
- Vieler, R., Humbert, F. & Marty, B. Evidence for a predominantly non-solar origin of nitrogen in the lunar regolith revealed by single grain analyses. *Earth Planet. Sci. Lett.* **167**, 47–60 (1999).
- Hashizume, K., Chaussidon, M., Marty, B. & Robert, F. Solar wind record on the moon: Deciphering presolar from planetary nitrogen. *Science* **290**, 1142–1145 (2000).
- Shizgal, B. D. & Arkos, G. G. Nonthermal escape of the atmospheres of Venus, Earth, and Mars. *Rev. Geophys.* **34**, 483–505 (1996).
- Seki, K., Elphic, R. C., Hirahara, M., Terasawa, T. & Mukai, T. On atmospheric loss of oxygen ions from Earth through magnetospheric processes. *Science* **291**, 1939–1941 (2001).
- Bochsler, P. Solar wind composition from the Moon. *Adv. Space Res.* **14**(6), 161–175 (1994).
- Mall, U., Christon, S., Kirsch, E. & Gloeckler, G. On the solar cycle dependence of the N^+/O^+ content in the magnetosphere and its relation to atomic N and O in the Earth's exosphere. *Geophys. Res. Lett.* **29**, doi:10.1029/2001GL013957 (2002).
- Hale, C. J. & Dunlop, D. Evidence for an early Archean geomagnetic field: a paleomagnetic study of the Komati Formation, Barberton Greenstone belt, South Africa. *Geophys. Res. Lett.* **11**, 97–100 (1984).
- Yoshihara, A. & Hamano, Y. Paleomagnetic constraints on the Archean geomagnetic field intensity obtained from komatiites of the Barberton and Belingwe greenstone belts, South Africa and Zimbabwe. *Precamb. Res.* **131**, 111–142 (2004).
- Kasprzak, W. T., Grebowsky, J. M. & Niemann, H. B. Superthermal > 36-eV ions observed in the near-tail region of Venus by the Pioneer Venus Orbiter Neutral Mass spectrometer. *J. Geophys. Res.* **96**, 11175–11187 (1991).
- Grebowsky, J. M., Kasprzak, W. T., Hartle, R. E., Mahajan, K. K. & Wagner, T. C. J. Superthermal ions detected in Venus dayside ionosheath, ionopause, and magnetic barrier regions. *J. Geophys. Res.* **98**, 9055–9064 (1993).

14. Abe, Y. Physical state of the very early Earth. *Lithos* **30**, 223–235 (1993).
15. Picone, J. M. et al. Enhanced empirical models of the thermosphere. *Phys. Chem. Earth C* **25**, 537–542 (2000).
16. Richards, P. G., Fennelly, J. A. & Torr, D. G. EUVAC: A solar EUV flux model for aeronomic calculations. *J. Geophys. Res.* **99**, 8981–8992 (1994).
17. Bills, B. G. & Ray, R. D. Lunar orbital evolution: A synthesis of recent results. *Geophys. Res. Lett.* **26**, 3045–3048 (1999).
18. Abe, M. & Ooe, M. Tidal history of the Earth–Moon dynamical system before Cambrian age. *J. Geodet. Soc. Jpn* **47**, 514–520 (2001).
19. Hashizume, K., Marty, B. & Wieler, R. Analyses of nitrogen and argon in single lunar grains: towards a quantification of the asteroidal contribution to planetary surface. *Earth Planet. Sci. Lett.* **202**, 201–216 (2002).
20. Heber, V. S., Baur, H. & Wieler, R. Helium in lunar samples analyzed by high-resolution stepwise etching: Implications for the temporal constancy of solar wind isotopic composition. *Astrophys. J.* **597**, 602–614 (2003).
21. Geiss, J. Solar wind composition and implications about the history of the solar system. *Proc. 13th Int. Cosmic Ray Conf.* 3375–3398 (Denver, University of Denver, 1973).
22. Mahaffy, P. R., Donahue, T. M., Atreya, S. K., Owen, T. C. & Niemann, H. B. Galileo probe measurements of D/H and $^3\text{He}/^4\text{He}$ in Jupiter's atmosphere. *Space Sci. Rev.* **84**, 252–263 (1998).
23. Busemann, H., Baur, H. & Wieler, R. Primordial noble gases in “phase Q” in carbonaceous and ordinary chondrites studied by closed-system stepped etching. *Meteorit. Planet. Sci.* **35**, 949–973 (2000).
24. Ozima, M. & Podosek, F. A. *Noble Gas Geochemistry* 2nd edn (Cambridge Univ. Press, Cambridge, 2002).
25. Hamano, Y. & Ozima, M. in *Terrestrial Rare Gases* (eds Alexander, E. C. Jr & Ozima, M.) 155–173 (Japan Scientific Societies Press, Tokyo, 1978).
26. Eugster, O., Terribini, D., Polnau, E. & Kramers, J. The antiquity indicator argon-40/argon-36 for lunar surface samples equilibrated by uranium-235-xenon-136 dating. *Meteorit. Planet. Sci.* **36**, 1097–1115 (2001).
27. Ozima, M., Miura, Y. N. & Podosek, F. A. Orphan radiogenic noble gases in lunar breccias: Evidence for planetary pollution of the Sun? *Icarus* **170**, 17–23 (2004).
28. Bernatowicz, T., et al. The regolith history of 14307. *Proc. Lunar Sci. Conf. VIII* 2763–2783, (1978).
29. Megrue, G. H. Spatial distribution of $^{40}\text{Ar}/^{36}\text{Ar}$ ages in lunar breccia 14301. *J. Geophys. Res.* **78**, 3216–3221 (1978).
30. Dalrymple, G. B. & Ryder, G. Argon-40/Argon-39 age spectra of Apollo 17 highlands breccia samples by laser step heating and the age of the Serenitatis basin. *J. Geophys. Res.* **101**, 26069–26084 (1996).
31. Murray, C. D. & Dermott, S. F. *Solar System Dynamics* (Cambridge Univ. Press, Cambridge, 1999).
32. Bilitza, D. International Reference Ionosphere 2000. *Radio Sci.* **36**, 261–275 (2001).
33. Wieler, R. in *Reviews in Mineralogy & Geochemistry* Vol. 47 (eds Porcelli, D., Ballentine, C. J. & Wieler, R.) (Geochemical Society, Mineralogical Society of America, Washington DC, 2002).

Acknowledgements We thank B. Marty, D. Stevenson and K. Zahnle for suggestions and comments that improved the manuscript. We owe much to the work of K. Hashizume, V. Heber, and co-workers, which inspired us to undertake this work. This work is supported by the 21st Century Center of Excellence (21CoE) SELIS (Dynamics of Sun–Earth–Life Interactive System) Program of Japan.

Author Contributions K.S. performed atmospheric escape estimation, and N.T. and H.S. ionospheric modelling.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.O. (EZZ03651@nifty.ne.jp).

ARTICLES

Oncogene-induced senescence as an initial barrier in lymphoma development

Melanie Braig¹, Soyoung Lee¹, Christoph Loddenkemper², Cornelia Rudolph³, Antoine H.F.M. Peters^{4,5}, Brigitte Schlegelberger³, Harald Stein², Bernd Dörken^{1,6}, Thomas Jenuwein⁵ & Clemens A. Schmitt^{1,6}

Acute induction of oncogenic Ras provokes cellular senescence involving the retinoblastoma (Rb) pathway, but the tumour suppressive potential of senescence *in vivo* remains elusive. Recently, Rb-mediated silencing of growth-promoting genes by heterochromatin formation associated with methylation of histone H3 lysine 9 (H3K9me) was identified as a critical feature of cellular senescence, which may depend on the histone methyltransferase Suv39h1. Here we show that Eμ-N-Ras transgenic mice harbouring targeted heterozygous lesions at the *Suv39h1*, or the *p53* locus for comparison, succumb to invasive T-cell lymphomas that lack expression of Suv39h1 or p53, respectively. By contrast, most N-Ras-transgenic wild-type ('control') animals develop a non-lymphoid neoplasia significantly later. Proliferation of primary lymphocytes is directly stalled by a Suv39h1-dependent, H3K9me-related senescent growth arrest in response to oncogenic Ras, thereby cancelling lymphomagenesis at an initial step. Suv39h1-deficient lymphoma cells grow rapidly but, unlike p53-deficient cells, remain highly susceptible to adriamycin-induced apoptosis. In contrast, only control, but not Suv39h1-deficient or p53-deficient, lymphomas senesce after drug therapy when apoptosis is blocked. These results identify H3K9me-mediated senescence as a novel Suv39h1-dependent tumour suppressor mechanism whose inactivation permits the formation of aggressive but apoptosis-competent lymphomas in response to oncogenic Ras.

In response to the activation of mitogenic oncogenes, checkpoint-mediated failsafe mechanisms such as apoptosis or cellular senescence may be recruited to terminate a pre-malignant condition before a fully transformed stage can develop^{1,2}. Premature senescence provoked by oncogenic Ras is associated with the accumulation of p53 and the D-type cyclin kinase inhibitor p16^{INK4a} that locks the retinoblastoma (Rb) protein in its active hypophosphorylated state³. Active Rb acts as a co-repressor of E2F transcription factors that transactivate genes important for S-phase progression⁴. Mapping of transcriptionally inactive sites in the vicinity of E2F-responsive promoters to heterochromatin foci enriched for H3K9me in senescent but not in quiescent cells supported the hypothesis that growth-promoting genes might be stably repressed in senescent cells by means of specific and possibly irreversible histone modifications⁵. The histone methyltransferase Suv39h1 is a candidate in this process, because it might methylate H3K9 as an Rb-bound enzyme in close proximity to E2F target genes^{6,7}, and H3K9me creates binding sites for HP1 proteins to form constitutive heterochromatin locally^{8,9}. Thus, on the basis of an inheritable principle of S-phase gene silencing¹⁰, H3K9me-mediated senescence might act as a tumour suppressor programme that limits the transforming capacity of oncogenic Ras.

Suv39h1 loss promotes Ras-driven lymphomagenesis

To determine the role of Suv39h1 defects in Ras-driven tumorigenesis *in vivo*, we intercrossed Eμ-N-Ras transgenic mice, constitutively expressing the activated N-RasG12D oncoprotein in the haematopoietic compartment, to *Suv39h1* knockout mice^{11,12}, and monitored offspring survival. Ras transgenic mice carrying no defective *Suv39h1* allele, hereafter referred to as controls, entered a terminal disease stage at a median age of 305 days. *Suv39h1*^{-/-} mice, comprising

Suv39h1^{-/-} females and, because of the X-linked *Suv39h1* locus, the equally Suv39h1-deficient *Suv39h1*^{-/y} males, succumbed to their condition significantly earlier (Fig. 1a; median time to death 66 days, $P = 0.0004$). Moreover, the time to death in *Suv39h1*^{+/-} mice was indistinguishable from the *Suv39h1*^{-/-} group, indicating insufficient Suv39h1 function in the underlying disease in heterozygous animals. Because p53-deficiency accelerates Ras-mediated tumour progression *in vivo*¹³ and disables Ras-induced senescence *in vitro*¹⁴, Eμ-N-Ras transgenic mice being wild-type or heterozygous for the *p53* locus were generated for comparison. Whereas *p53*^{+/-} mice behaved in the same way as controls, *p53*^{+/-} mice became terminally ill almost as fast as mice lacking one or both *Suv39h1* alleles (Fig. 1a; median time to death 141 days for *p53*^{+/-}, $P = 0.02$ compared with controls). No overt signs of disease were observed during more than a year in non-transgenic mice lacking one *p53* or at least one *Suv39h1* allele. Thus, allelic defects at the *Suv39h1* or the *p53* locus markedly accelerate disease manifestation in the context of activated Ras.

At necropsy, mice that died early (that is, 180 days or less after birth; 'early death group') invariably displayed thymic and splenic but not hepatic enlargements due to a clonal lymphoblastic T-cell infiltration accompanied by a massive leukaemic burden and aggressive invasion into non-lymphoid visceral organs including the lungs. In stark contrast, necropsy of mice that died later (that is, more than 180 days after birth; 'late death group') revealed a distinct gross pathology with massively enlarged livers and spleens, but only occasional signs of thymic or visceral organ involvement (Fig. 1b, Supplementary Table 1 and Supplementary Fig. S1a, b). Histological examination, immunohistochemistry and immunophenotyping confirmed a myeloid origin of this neoplastic entity in accordance with histiocytic sarcomas previously described in the Eμ-N-Ras transgenic mouse model¹¹ (Fig. 1c and Supplementary Fig. S1c),

¹Charité – Universitätsmedizin Berlin/Haematology-Oncology, 13353 Berlin, Germany. ²Charité – Universitätsmedizin Berlin/Department of Pathology, 12200 Berlin, Germany.

³Institute of Cell and Molecular Pathology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany. ⁴Friedrich Miescher Institute for Biomedical Research, 4058 Basel, Switzerland. ⁵Research Institute of Molecular Pathology, 1030 Vienna, Austria. ⁶Max-Delbrück-Center for Molecular Medicine, 13125 Berlin, Germany.

discriminating it from an aggressive T-cell malignancy (Fig. 1c and Supplementary Fig. S1d, e). Altered lymphoid differentiation in the absence of Suv39h1 could be ruled out by a comprehensive immunophenotypic analysis of the cellular compositions in various haematopoietic compartments (Supplementary Fig. S2 and Supplementary Table 2). Taken together, these results show that *Suv39h1* or *p53* defects disable a tumour-suppressive programme that protects against Ras-driven lymphomagenesis.

We next investigated whether oncogenic Ras might select against *p53* or *Suv39h1* expression in lymphomas that formed in *p53*^{+/-} or *Suv39h1*^{+/-} mice, respectively. According to time-to-death data implying that Suv39h1 and p53 limit Ras-induced lymphomagenesis, Suv39h1 transcripts were undetectable in all lymphomas derived from *Suv39h1*^{+/-} mice (Fig. 2a; 12 of 12 tumours tested), and Eμ-*N-Ras* lymphomas arising in *p53*^{+/-} animals invariably lost the remaining wild-type *p53* allele (Fig. 2a; four of five tumours tested), rendering them *Suv39h1*-null (hereafter referring to all

Suv39h1-deficient lymphomas) or *p53*-null, respectively. Unlike *p53* loss by allelic deletion, lack of Suv39h1 expression probably reflects a selective advantage for progenitor cells in which the remaining *Suv39h1* allele maps to the inactivated X chromosome¹⁵, because we did not detect gross genomic deletions at this locus in any *Suv39h1*^{+/-}-derived lymphoma by allele-specific polymerase chain reaction (PCR) (data not shown). Thus, Ras-driven lymphomagenesis is promoted by selective inactivation of Suv39h1 or *p53* function.

The rapid formation of *Suv39h1*-null or *p53*-null lymphomas raises the possibility that they might select for high-level expression of oncogenic Ras. However, immunoblot analysis confirmed remark-

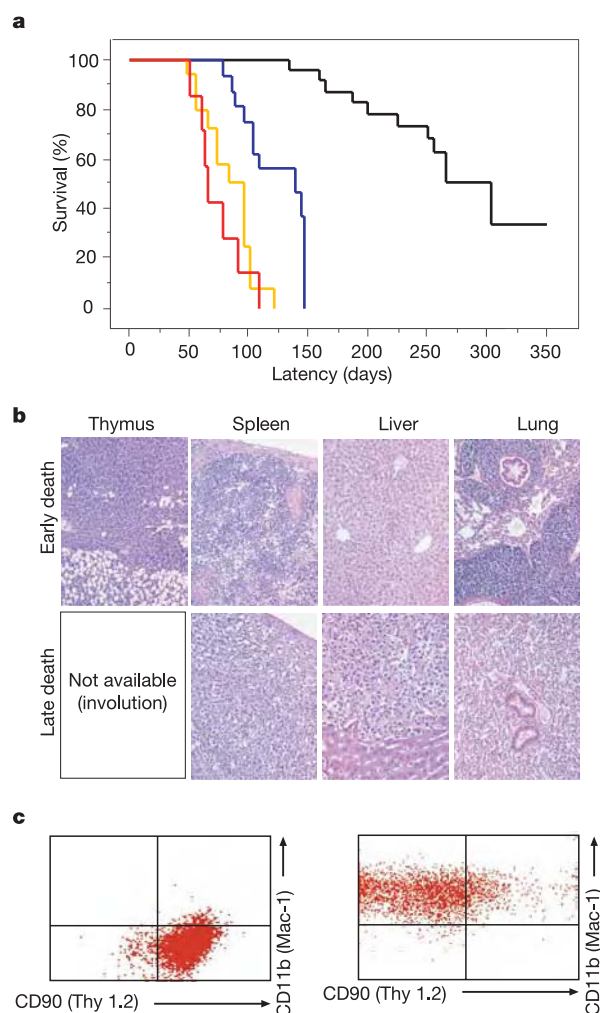


Figure 1 | *Suv39h1*- and *p53*-defects license aggressive lymphomas in response to oncogenic Ras. **a**, Kaplan-Meier plot showing time-to-death latencies in Eμ-*N-Ras* transgenic *Suv39h1*^{-/-} females and *Suv39h1*^{-/-} males (*n* = 7; red), *Suv39h1*^{+/-} (*n* = 17; orange), *p53*^{+/-} (*n* = 19; blue), and pooled control mice (*n* = 63; black). **b**, Representative photomicrographs of tissue sections at terminal disease stages of the indicated organs derived from early-death or late-death group animals (shown are *Suv39h1*-inactivated (top) and control cases (bottom)). Staining was with haematoxylin and eosin. **c**, Immunophenotyping of early-death (*Suv39h1*-inactivated, left) versus late-death (control, right) neoplastic cell populations by two-colour flow cytometry for a T-cell versus a myeloid marker antigen (CD90 versus CD11b).

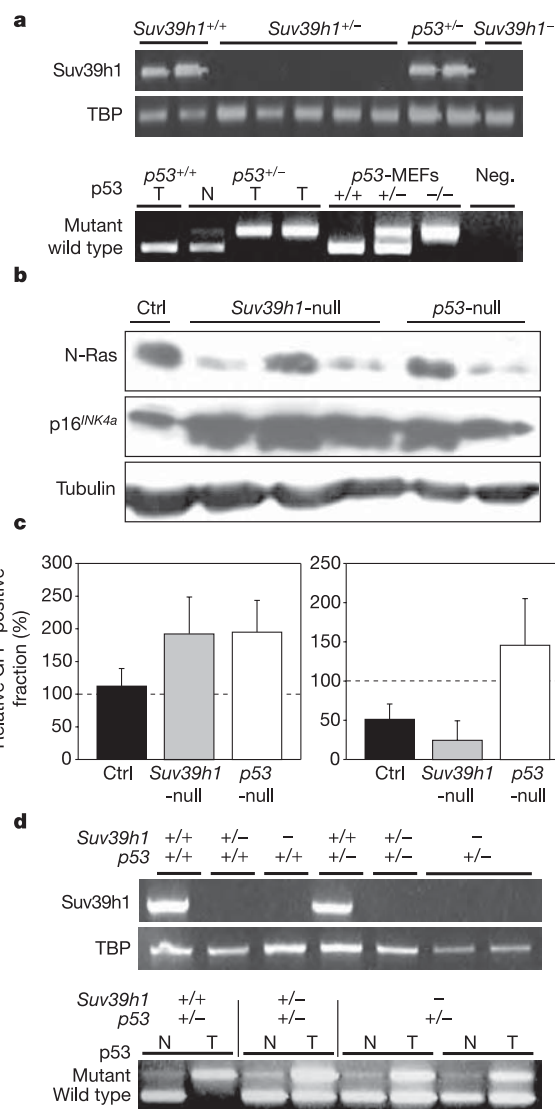


Figure 2 | Oncogenic Ras selects against *Suv39h1*, ARF or *p53* in primary lymphomas. **a**, Suv39h1 RT-PCR (top; TBP, internal control) and allele-specific *p53* genomic PCR to detect the wild-type or knockout (mutant) allele in tumour (T) and corresponding normal (N) tissues (bottom; mouse embryo fibroblast (MEF) DNA, internal control). **b**, Immunoblot analyses of control (ctrl), *Suv39h1*-null and *p53*-null lymphomas for N-Ras, p16^{INK4a} and α-tubulin (loading control). **c**, Relative changes of the GFP-positive fraction 48 h after infection with MSCV-p16^{INK4a}-IRES-GFP (left) or MSCV-ARF-IRES-GFP (right) in control (*n* = 3), *Suv39h1*-null (*n* = 8) and *p53*-null lymphomas (*n* = 5) determined by flow cytometry; error bars denote s.d. **d**, Suv39h1-RT-PCR (top) and *p53* PCR (bottom; as in **a**) in lymphomas derived from *Suv39h1*^{+/-}; *p53*^{+/-} and *Suv39h1*^{-/-}; *p53*^{+/-} mice.

able inter-sample variability of Ras protein levels within any given genotype, but having similar ranges between different genotypes (Fig. 2b). Particularly high p16^{INK4a} protein levels, induced by the Ras–Raf–MAP kinase/ERK kinase (MEK) signalling cascade¹⁶, were found in *Suv39h1*-null cells, underscoring the putative role of p16^{INK4a} acting upstream of Suv39h1 (Fig. 2b and Supplementary Fig. S3a). To dissect oncogenic signalling in *Suv39h1*-null or *p53*-null lymphoma cells further, we introduced complementary DNAs

encoding the alternative reading frame product of the *INK4a* locus ARF or p16^{INK4a} into these cells and monitored whether a particular gene activity was enriched for or selected against by means of co-expressed green fluorescent protein (GFP). Consistent with high endogenous p16^{INK4a} protein levels (Fig. 2b and Supplementary Fig. S3b) was our observation that *Suv39h1*-null and *p53*-null lymphoma cells were enriched for the *INK4a* moiety within a 48-h period, whereas control lymphomas maintained stable p16^{INK4a} levels as indicated by an unchanged GFP fraction (Fig. 2c). Both control and *Suv39h1*-null lymphomas did not tolerate above-physiological levels of ARF, whereas *p53*-null lymphomas, consistent with ARF's acting upstream of p53, accepted exogenous expression of this moiety. Accordingly, endogenous ARF protein expression was low in *Suv39h1*-null lymphomas *in situ* compared with high levels in *p53*-null lymphomas and was undetectable in control lymphomas; both *Suv39h1*-null and control lymphomas retained wild-type *p53* (five of five cases per genotype sequenced; Supplementary Fig. S3b). Both *Suv39h1*[−] and *Suv39h1*^{+/−} Eμ-*N-Ras* mice being heterozygous for *p53* succumbed—within the time-to-death range known from *Suv39h1*-inactivated animals—to lymphomas lacking Suv39h1 expression but retaining the remaining *p53* wild-type allele (Fig. 2d; three of three tumours tested). Hence, *Suv39h1* inactivation alleviates the pressure to co-inactivate p53 in response to oncogenic Ras during lymphoma development, although *Suv39h1*-null lymphomas remain highly sensitive to exogenous p53-activating triggers such as ARF overexpression.

Chromosomal stability and Suv39h1 deficiency

Because B-cell lymphomas that formed in a non-transgenic context after long latency in the absence of *Suv39h1* and the closely related *Suv39h2* genes display hyperdiploid karyotypes and to some extent so-called 'butterfly' chromosomes reminiscent of a chromosomal segregation defect¹², we next analysed the karyotypes and Suv39h2 expression in Ras-driven T-cell lymphomas retaining or losing Suv39h1 expression. By scanning a large number of metaphases, we observed similarly modest deviations from the normal number of 40 chromosomes in control and *Suv39h1*-null lymphomas (40.3 ± 0.7 and 40.1 ± 1.2 chromosomes, respectively; errors are $\pm$ s.d.) and a tendency towards hyperdiploidy in *p53*-null lymphomas (43.7 ± 2.2 chromosomes). In particular, no non-segregated 'butterfly' chromosomes were detected in any metaphase. Suv39h2 is expressed in Ras-lymphomas throughout the genotypes, including *Suv39h1*-null cases (three of three samples tested; Supplementary Fig. S3c). More refined molecular cytogenetic analyses by spectral karyotyping underscored the numeric ranges reported, and confirmed that *Suv39h1*-null lymphomas display either no overt chromosomal aberrations or similar ones to those detected in control lymphomas (Fig. 3a and Supplementary Table 3). More precisely, gains of chromosomes 10 and 17 as well as chromosome 15 translocations were identified in subsets of both genotypes, possibly indicating a clonal evolution contributing to lymphoma progression. Thus, Suv39h1 deficiency promotes lymphomagenesis in response to oncogenic Ras without mis-segregation-driven chromosomal instability.

Availability of drug-inducible apoptosis and senescence

Recapitulating their aggressive behaviour and equally high proliferative capacity *in vivo* (Fig. 1b and Supplementary Fig. S3d, e), isolated control, *Suv39h1*-null and *p53*-null *Ras* transgenic lymphoma cells could be readily established as primary cultures *in vitro*, where they all grew exponentially (Fig. 3b). Whereas *Suv39h1*-null lymphomas displayed a slight growth advantage at best in comparison with controls, cell numbers in the *p53*-null group increased most rapidly, raising the question of whether an apoptotic defect in the absence of p53 might contribute to this phenotype (as revealed for spontaneous apoptosis *in situ*; Supplementary Fig. S3f). To examine the apoptotic capacity of Ras-driven lymphomas *in vitro*,

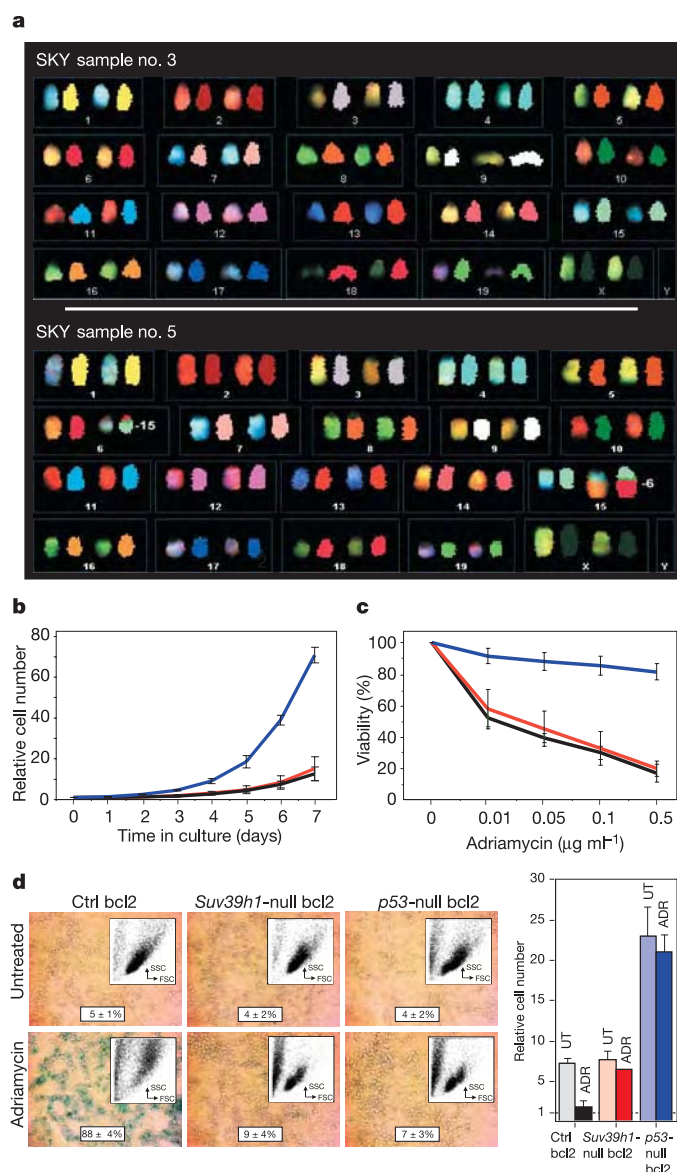


Figure 3 | Suv39h1 loss produces chromosomally stable senescence-defective, but apoptosis-competent, Ras lymphomas. a, Spectral karyotypes of individual *Suv39h1*-null lymphomas; sample no. 3 with a normal karyotype, and sample no. 5 with a T(6;15) translocation (see Supplementary Table 3 for details). **b**, Growth curve analysis of freshly isolated control (n = 5; black), *Suv39h1*-null (n = 14; red) and *p53*-null lymphomas (n = 4; blue). **c**, Twenty-four-hour viability after exposure to adriamycin; cells as in **b**. **d**, Left: SA-β-galactosidase activity in *bcl2*-infected lymphomas 5 days after no treatment or 0.1 μg ml⁻¹ adriamycin (n = 3; percentages reflect positive cells (means $\pm$ s.d.)). Insets show, after 3 days, increased cell size (forward scatter, FSC) and granularity (side scatter, SSC) reminiscent of cellular senescence in treated control cells only. Right: growth curve analysis (n = 4); only control cells arrest after treatment with adriamycin (ADR; *P* = 0.0001). UT, untreated. Relative cell numbers are the proportion of viable cells detectable after 5 days.

primary lymphoma cells of the different genotypes were exposed to the DNA-damaging agent adriamycin in a short-term cytotoxicity assay. Whereas *p53*-null lymphomas were highly resistant to adriamycin-induced cell death as expected, *Suv39h1*-null and control cells underwent massive apoptosis even at low drug concentrations (Fig. 3c; the apoptotic nature of cell death was confirmed by staining with Annexin V; data not shown). Because anticancer drugs can acutely provoke a terminal growth arrest that exhibits morphological, biochemical and genetic features indistinguishable from premature senescence in response to oncogenic Ras^{3,17,18}, we examined drug-inducible senescence in Ras-driven lymphomas that were protected from cell death by expression of the anti-apoptotic regulator Bcl2. Adriamycin produced a tight growth arrest accompanied by strongly induced senescence-associated

β -galactosidase (SA- β -gal) activity¹⁹ and morphological changes typical for cellular senescence in control lymphomas, whereas both *Suv39h1*-null and *p53*-null cells continued to proliferate and remained virtually negative for SA- β -gal activity (Fig. 3d). Moreover, control lymphomas responded to the drug with a sharp decline in 5-bromo-2-deoxyuridine incorporation paralleled by an almost complete accumulation in the G1 fraction, whereas *Suv39h1*-null cells displayed little change in G1 and S fractions after treatment (data not shown). Irrespective of their ability to enter senescence, lymphomas of all genotypes contained higher amounts of p16^{INK4a}, and *Suv39h1*-null and *p53*-null lymphomas induced ARF protein concentrations to various extents in response to adriamycin (compare Supplementary Fig. S3g with Supplementary Fig. S3b). Taken together, these results show that loss of *Suv39h1* expression renders Ras-driven lymphomas senescence-incompetent without compromising their apoptotic capacity.

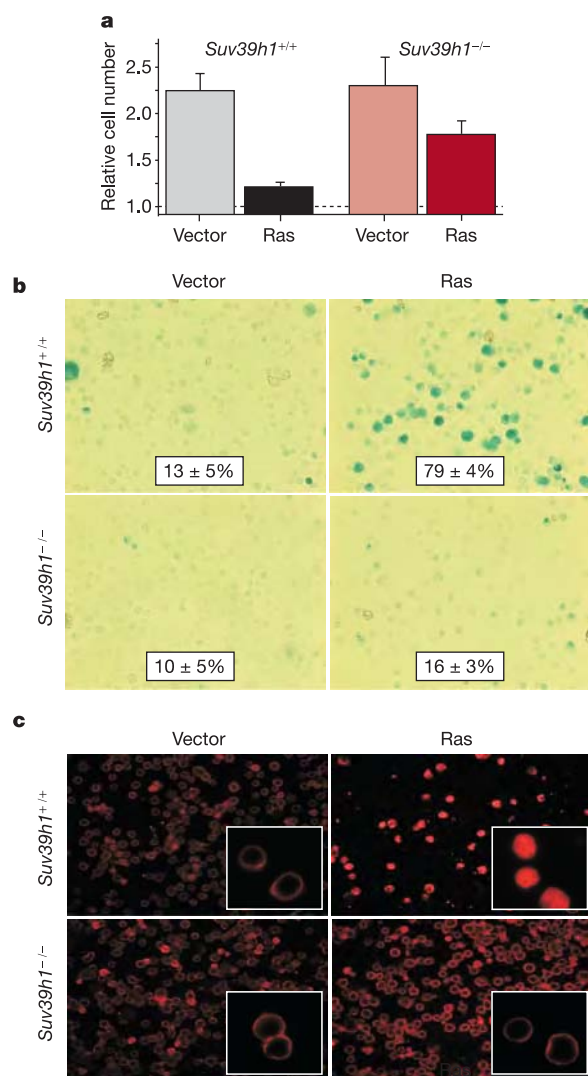


Figure 4 | Suv39h1/H3K9me-controlled senescence limits Ras-mediated oncogenicity. Primary *Suv39h1*^{+/+} and *Suv39h1*^{-/-} splenocytes were stably infected with oncogenic *ras* or an empty vector as control, and were plated at identical cell numbers immediately on completion of antibiotic selection. **a**, Growth curve analysis at day 6 ($n = 4$ independent preparations each); error bars denote s.d. Relative cell numbers are the proportion of viable cells detectable after 6 days. **b**, **c**, Cytospin preparations of the day-6 samples assayed for SA- β -galactosidase activity ($n = 3$; percentages reflect positive cells (means $\pm$ s.d.); note that cells may gradually enter an SA- β -galactosidase-positive state) (**b**), and stained by immunofluorescence for trimethylated H3K9 (medium power-field) and HP1- γ (high-power inset) (**c**).

Suv39h1 dependence of Ras-provoked senescence

To determine whether *Suv39h1*-controlled cellular senescence might directly limit Ras-induced transformation in the lymphoid compartment, a retroviral construct encoding activated Ras was introduced into primary *Suv39h1*^{+/+} or *Suv39h1*^{-/-} splenocytes. The exogenous expression of oncogenic Ras in *Suv39h1*^{+/+} lymphocytes virtually stalled their net proliferative capacity but allowed the *Suv39h1*^{-/-} population to increase significantly (Fig. 4a, $P = 0.0002$). There was no difference in the growth potential between vector-infected *Suv39h1*^{+/+} and *Suv39h1*^{-/-} lymphocytes. Accordingly, only Ras-expressing *Suv39h1*^{+/+} lymphocytes displayed a strong SA- β -gal activity 6 days after Ras induction throughout the population, whereas very little SA- β -gal activity was sporadically detectable in the absence of either or both oncogenic Ras and *Suv39h1* (Fig. 4b). In response to Ras only *Suv39h1*^{+/+} cells displayed a sharp increase in H3K9 trimethylation and produced an intense yet grainy staining pattern for HP1- γ , possibly indicating a murine equivalent of Ras-dependent senescence-associated chromatin foci so far described for human cells only⁵ (Fig. 4c). Thus, *Suv39h1*-mediated H3K9me-associated senescence serves as an early barrier to mitogenic Ras signalling in primary lymphocytes.

Because acetylated H3K9 cannot be methylated by *Suv39h1* (ref. 6), and because *Suv39h1* might locally reinforce transcriptional repression through association with DNA methyltransferases^{20,21}, we reasoned that a combined treatment of a histone deacetylase inhibitor such as trichostatin A (TSA) with the DNA-demethylating agent 2-deoxy-5-azacytidine (DAC) might at least partly mimic *Suv39h1*

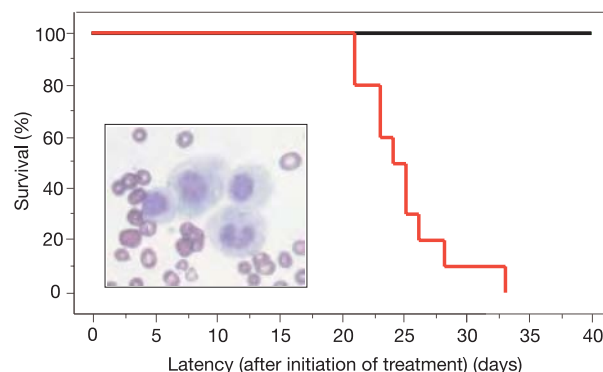


Figure 5 | Methylation-targeting therapy mimics *Suv39h1* loss in a Ras transgenic context *in vivo*. Time-to-death observation in *Eμ-N-Ras* transgenic mice repeatedly treated with trichostatin A (TSA) and 2-deoxy-5-azacytidine (DAC) or vehicle only (mock). Latency indicates the time after initiation (at about 6 weeks of age) of the 4-week protocol until a terminal disease stage was entered or unexpected death occurred (mock-treated, $n = 6$, black; TSA/DAC-treated (developing leukaemic T-cell lymphomas, inset), $n = 11$, red).

inactivation. Whereas $\text{E}\mu\text{-N-Ras}$ mice remained free of disease symptoms for more than 8 weeks when exposed to the vehicle only, they all developed clonal CD3^+ T-cell lymphomas with leukaemic dissemination—similar, although somewhat more pleomorphic and mature, to *Suv39h1*-null Ras lymphomas—within 5 weeks under the TSA/DAC protocol (Fig. 5 and Supplementary Fig. S4a, b). Although the combination therapy at the dosage and schedule chosen seemed to be toxic in itself, lethal side effects observed in some of the wild-type mice under the TSA/DAC protocol occurred significantly later ($P = 0.0358$), with histopathological aspects of myelosuppression and acute hepatic toxicity, but no evidence for lymphoma at necropsy. Thus, pharmacological targeting *in vivo* of epigenetic modifications including Suv39h1-governed histone and DNA methylation resembles the long-term outcome observed when *Suv39h1* is genetically inactivated. The tumour-promoting action of methylation-targeting compounds in this genetic context sharply contrasts with their clinically promising use in myelodysplastic syndrome and some acute myeloid leukaemias²², highlighting the complexity of epigenetic alterations in different neoplastic and pre-neoplastic scenarios. Thus, these results provide evidence that Suv39h1-mediated histone modification controls cellular senescence as a tumour-suppressive programme that both directly limits transformation in response to oncogenic Ras *in vitro* and abrogates Ras-driven lymphoma formation *in vivo*.

Relevance of oncogene-induced senescence *in vivo*

Cellular senescence provides a key barrier to oncogene-mediated transformation *in vitro*, but its role in tumour suppression *in vivo* is poorly demonstrated^{1,3,23}. Our results have now identified H3K9me-related senescence as a mechanism that efficiently cancels Ras-induced lymphomagenesis *in vivo*, promoting aggressive lymphomas in all $\text{E}\mu\text{-N-Ras}$ mice lacking at least one *Suv39h1* allele, whereas most transgenic control animals remain free of lymphoma and succumb to non-lymphoid neoplasms much later. Although the translocation breakpoint on chromosome 15 detected in several cases by spectral karyotyping may involve the *c-myc* locus, alternative alterations sufficient to bypass Ras-induced senescence were found in virtually all lymphomas tested. Ras-provoked senescence is a directly Suv39h1-dependent tumour suppressor mechanism controlled by both ARF and p53 that acts as the prime barrier against lymphoma formation. Accordingly, no lymphoma formed without an expression defect in one of these senescence regulators. Premature senescence has been demonstrated mostly in response to above-physiological expression levels of oncogenic H-Ras *in vitro*. In contrast, oncogenic K-Ras expressed from its endogenous promoter has been reported to promote proliferation directly without causing cellular senescence in certain circumstances *in vitro*, but its immortalizing and transforming capacity *in vivo* is highly dependent on context and cell type, and seems to rely on a titrated attenuation of the MAP kinase pathway, whose activity is critical for the induction of cellular senescence^{14,24,25}. The implication of Suv39h1 function in Ras-provoked cellular responses might also depend on the cell type, because mouse embryo fibroblasts lacking *Suv39h1/2* genes still arrest in response to oncogenic Ras⁵ and because we did not observe accelerated simultaneous manifestation of histiocytic sarcomas at the time that *Suv39h1*^{-/-} or *p53*^{+/-} mice succumbed to their lymphomas.

Although progression models for many solid cancers from pre-malignant tumours to full-blown malignancies have been established, no such precursor lesion has yet been identified in the process of lymphoma development, and it remains technically challenging to detect a putative growth-arrested lesion formed at an unknown level anywhere in the lympho-haematopoietic compartment. Although no *Suv39h1* alterations have been reported in manifest human cancers so far, deregulation of Suv39h1 function might also be attributed to defective binding partners, because naturally occurring Rb mutants fail to complex with Suv39h1, thereby disrupting its spatial action on

E2F-responsive promoters⁷. Moreover, another Rb-binding H3K9 methyltransferase, Riz1, is inactivated by mutations in human cancers²⁶, and Riz1-deficient mice are lymphoma-prone²⁷, possibly because of a similar H3K9me-related senescence defect.

Our results also have important ramifications for therapeutic strategies, because the precise genetic defect ultimately licensing Ras-driven lymphoma formation might interfere with both drug-inducible apoptosis and senescence programmes. Although *Suv39h1*-null lymphomas are phenotypically similar to those lacking p53 with regard to their early onset, they present with a sensitized p53 machinery susceptible to exogenous apoptotic triggers such as ARF overexpression or DNA-damaging drugs. Interestingly, control lymphomas that formed in response to Ras without explicit defects in Suv39h1 or p53 can bypass oncogene-induced senescence through the loss or critical reduction of ARF expression, but may still recall a senescence programme when triggered by exogenous stimuli such as DNA-damaging anticancer drugs. Because both drug-inducible apoptosis and senescence contribute to the outcome of cancer therapy¹⁸, we propose that the identification and functional assessment of genetic defects inactivating oncogene-provoked failsafe mechanisms will be important for optimizing treatment strategies.

METHODS

Mouse strains and tumour monitoring. $\text{E}\mu\text{-N-Ras}$ transgenic mice were intercrossed to mice harbouring targeted deletions in the *Suv39h1* or *p53* locus. All strains were backcrossed into a C57BL/6 strain background. Genotyping was performed as described²⁸. Time to death was defined as the latency between birth and unexpected death or a terminal disease stage as indicated by more than 20% weight loss or other symptoms of severe sickness. Statistical analysis of Kaplan–Meier survival plots is based on the log-rank (Mantel–Cox) test. After killing of the mice with CO_2 , single-cell suspensions were isolated from enlarged organs, and tissues were processed as described for histopathology and subsequent staining with haematoxylin and eosin^{18,28}. Blood smears were taken as described²⁸. Spleens from 8–12-week-old non-transgenic mice served as the source for primary splenocyte preparations. For combined TSA (Sigma) and DAC (Sigma) treatments *in vivo*, 39–45-day-old mice received TSA (0.5 mg kg^{-1} body weight) subcutaneously once a day and DAC (2 mg kg^{-1} body weight) intraperitoneally three times weekly or mock treatment (solvents only) for up to 4 weeks, and were monitored daily for signs of sickness.

Cell culture, gene transfer and analysis of cell growth and integrity. Isolated splenocytes and lymphoma cells were cultured on irradiated NIH 3T3 feeder cells as described²⁸. Numeric karyotypic analysis was performed on more than 200 4',6-diamidino-2-phenylindole (DAPI)-stained metaphase spreads of at least three lymphoma samples per genotype²⁹, and spectral karyotyping (SKY) was performed as described³⁰. T-cell clonality was assessed by multiplex PCR amplification of genomic DNA isolated from short-term-cultured lymphoma cells, using a panel of V(D)J primers for the T-cell receptor (TCR) β -chain³¹. Retroviral gene transfer was performed as described²⁹; all retroviruses were MSCV-based constructs that either co-express GFP (namely MSCV-p16^{INK4a}-IRES-GFP; MSCV-ARF-IRES-GFP) or a selectable antibiotic resistance gene (namely MSCV-bcl2-puro; MSCV-H-RasV12-puro; MSCV-Suv39h1-hygro), including the corresponding vector-only controls. For enrichment experiments based on GFP co-expressing vectors, freshly infected lymphoma cells were adjusted to about 10% GFP-positive cells by admixing uninfected cells of the same lymphoma population, and the fraction of GFP-positive cells was reassessed by flow cytometry after 48 h in culture. Cell viability and cell numbers were analysed by Trypan blue dye exclusion; 5-bromo-2-deoxyuridine incorporation and cell-cycle analyses were performed as described²⁹. Spontaneous apoptosis *in situ* was measured as described²⁸ and was quantified by integrated scanning of fluorescence photomicrographs using the NIH image 1.63 software. Apoptosis induced by the DNA-damaging compound adriamycin (Sigma) was quantified by staining with annexin V in accordance with the manufacturer's protocol (BD Pharmingen). Comparisons of means and standard deviations were performed with the unpaired *t*-test. Senescence-associated β -galactosidase activity was assessed in cytosin preparations as described¹⁸; senescence-associated changes in the cellular morphology were quantified by flow-cytometric light-scatter analysis. All experiments were performed with reproducible results in at least duplicate.

Gene sequence and expression analysis. Immunoblotting was performed with whole-cell lysates resolved by 15% SDS-PAGE and transferred to an Immobilon-P membrane (Millipore) by using antibodies against ARF (ab80; Abcam; 1:1000

dilution), Bcl2 (no. 554087; BD Pharmingen; 1:2000), H-Ras (OP23; Oncogene; 1:1000), N-Ras (OP25; Oncogene; 1:500) and p16^{INK4a} (M-156; Santa Cruz; 1:500), with antibodies against α -tubulin (T5168; Sigma; 1:8000) as a loading control²⁸. To perform immunohistochemistry on formalin-fixed, paraffin-embedded tissue sections and immunocytochemistry on formalin-fixed cytopsin preparations for p16^{INK4a} (F-12; Santa Cruz; 1:2000), ARF (ab80; Abcam; 1:20), Ki67 (Tec-3; DakoCytomation; 1:2000), CD3 (no. 1580; Dako; 1:100), CD11b (M1/70; BD Pharmingen; 1:50) and terminal deoxynucleotidyl transferase (DakoCytomation; 1:20), slides were immersed in sodium citrate buffer solutions at pH 6.0, heated in a high-pressure cooker for 5 min, rinsed in cold water and washed in Tris-buffered saline (pH 7.4), and were then treated with a peroxidase-blocking reagent (DakoCytomation) before incubation for 1 h with the respective primary antibody. Binding was detected by the EnVision peroxidase kit using diaminobenzidine as the chromogenic substrate (K 4010; DakoCytomation). For immunofluorescence, cells were fixed in 2% paraformaldehyde, permeabilized with 0.2% Triton X-100 in PBS, preincubated in rabbit normal serum and incubated with primary antibodies against HP1- γ (no. 07-332; Upstate; 1:1000), and trimethylated H3K9 (ab8898; Abcam; 1:1000), followed by the secondary anti-rabbit IgG antibody (Alexa Fluor 594; Molecular Probes; 1:2000). For immunophenotyping by flow cytometry, directly fluorescence-conjugated antibodies against B220 (CD45R), surface IgM, Thy1.2 (CD90), TCR- β , TCR- γ/δ , GR-1, NK1.1, CD3, CD4, CD5, CD8, CD19, CD43, CD68, CD86, CD117, CD138 (all BD Pharmingen) and Mac-1 (CD11b; Acris; 1:400) were used as described²⁸.

Suv39h1 and Suv39h2 transcript expression levels, compared with TATA-box-binding-protein (TBP) expression as an internal control, were analysed after gene-specific reverse transcription (Superscript; Invitrogen) and PCR amplification of lymphoma cell total RNA.

RT-PCR products of exons 4–8 of the p53 gene were sequenced as described²⁸.

Received 19 January; accepted 26 May 2005.

- Campisi, J. Cellular senescence as a tumour-suppressor mechanism. *Trends Cell Biol.* **11**, S27–S31 (2001).
- Schmitt, C. A. Senescence, apoptosis and therapy—cutting the lifelines of cancer. *Nature Rev. Cancer* **3**, 286–295 (2003).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88**, 593–602 (1997).
- Trimarchi, J. M. & Lees, J. A. Sibling rivalry in the E2F family. *Nature Rev. Mol. Cell Biol.* **3**, 11–20 (2002).
- Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).
- Rea, S. *et al.* Regulation of chromatin structure by site-specific histone H3 methyltransferases. *Nature* **406**, 593–599 (2000).
- Nielsen, S. J. *et al.* Rb targets histone H3 methylation and HP1 to promoters. *Nature* **412**, 561–565 (2001).
- Bannister, A. J. *et al.* Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature* **410**, 120–124 (2001).
- Lachner, M., O'Carroll, D., Rea, S., Mechtler, K. & Jenuwein, T. Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins. *Nature* **410**, 116–120 (2001).
- Ait-Si-Ali, S. *et al.* A Suv39h-dependent mechanism for silencing S-phase genes in differentiating but not in cycling cells. *EMBO J.* **23**, 605–615 (2004).
- Haupt, Y., Harris, A. W. & Adams, J. M. Retroviral infection accelerates T lymphomagenesis in E mu-N-ras transgenic mice by activating c-myc or N-myc. *Oncogene* **7**, 981–986 (1992).
- Peters, A. H. *et al.* Loss of the Suv39h histone methyltransferases impairs mammalian heterochromatin and genome stability. *Cell* **107**, 323–337 (2001).
- Kemp, C. J., Donehower, L. A., Bradley, A. & Balmain, A. Reduction of p53 gene dosage does not increase initiation or promotion but enhances malignant progression of chemically induced skin tumors. *Cell* **74**, 813–822 (1993).
- Lin, A. W. *et al.* Premature senescence involving p53 and p16 is activated in response to constitutive MEK/MAPK mitogenic signalling. *Genes Dev.* **12**, 3008–3019 (1998).
- Lyon, M. F. Gene action in the X-chromosome of the mouse (*Mus musculus* L.). *Nature* **190**, 372–373 (1961).
- Ohtani, N. *et al.* Opposing effects of Ets and Id proteins on p16^{INK4a} expression during cellular senescence. *Nature* **409**, 1067–1070 (2001).
- Chang, B. D. *et al.* A senescence-like phenotype distinguishes tumour cells that undergo terminal proliferation arrest after exposure to anticancer agents. *Cancer Res.* **59**, 3761–3767 (1999).
- Schmitt, C. A. *et al.* A senescence program controlled by p53 and p16^{INK4a} contributes to the outcome of cancer therapy. *Cell* **109**, 335–346 (2002).
- Dimri, G. P. *et al.* A biomarker that identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
- Nguyen, C. T. *et al.* Histone H3-lysine 9 methylation is associated with aberrant gene silencing in cancer cells and is rapidly reversed by 5-aza-2'-deoxycytidine. *Cancer Res.* **62**, 6456–6461 (2002).
- Fuks, F., Hurd, P. J., Deplus, R. & Kouzarides, T. The DNA methyltransferases associate with HP1 and the SUV39H1 histone methyltransferase. *Nucleic Acids Res.* **31**, 2305–2312 (2003).
- Leone, G., Voso, M. T., Teofili, L. & Lubbert, M. Inhibitors of DNA methylation in the treatment of hematological malignancies and MDS. *Clin. Immunol.* **109**, 89–102 (2003).
- Rane, S. G., Cosenza, S. C., Mettut, R. V. & Reddy, E. P. Germ line transmission of the Cdk4(R24C) mutation facilitates tumorigenesis and escape from cellular senescence. *Mol. Cell Biol.* **22**, 644–656 (2002).
- Guerra, C. *et al.* Tumor induction by an endogenous K-ras oncogene is highly dependent on cellular context. *Cancer Cell* **4**, 111–120 (2003).
- Tuveson, D. A. *et al.* Endogenous oncogenic K-ras(G12D) stimulates proliferation and widespread neoplastic and developmental defects. *Cancer Cell* **5**, 375–387 (2004).
- Kim, K. C., Geng, L. & Huang, S. Inactivation of a histone methyltransferase by mutations in human cancers. *Cancer Res.* **63**, 7619–7623 (2003).
- Steele-Perkins, G. *et al.* Tumor formation and inactivation of RIZ1, an Rb-binding member of a nuclear protein-methyltransferase superfamily. *Genes Dev.* **15**, 2250–2262 (2001).
- Schmitt, C. A., McCurrach, M. E., de Stanchina, E., Wallace-Brodeur, R. R. & Lowe, S. W. INK4a/ARF mutations accelerate lymphomagenesis and promote chemoresistance by disabling p53. *Genes Dev.* **13**, 2670–2677 (1999).
- Schmitt, C. A. *et al.* Dissecting p53 tumour suppressor functions *in vivo*. *Cancer Cell* **1**, 289–298 (2002).
- Frank, O. *et al.* Tumor cells escape suicide gene therapy by genetic and epigenetic instability. *Blood* **104**, 3543–3549 (2004).
- Hamrouni, A., Aublin, A., Guillaume, P. & Maryanski, J. L. T cell receptor gene rearrangement lineage analysis reveals clues for the origin of highly restricted antigen-specific repertoires. *J. Exp. Med.* **197**, 601–614 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank A. Harris, T. Jacks and M. Serrano for mice; S. W. Lowe for retroviral constructs; B. Teichmann and S. Spieckermann for technical assistance; and A. Lee, M. Reimann and P. Kahlem for discussions and editorial advice. This work was supported by grants from the European Union (to B.S.) and from the Deutsche Krebshilfe (to C.A.S.). S.L. is a postdoctoral fellow of the José Carreras Leukemia Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.A.S. (clemens.schmitt@charite.de).

LETTERS

The obscuration by dust of most of the growth of supermassive black holes

Alejo Martínez-Sansigre¹, Steve Rawlings¹, Mark Lacy², Dario Fadda², Francine R. Marleau², Chris Simpson³, Chris J. Willott⁴ & Matt J. Jarvis¹

Supermassive black holes underwent periods of exponential growth during which we see them as quasars in the distant Universe. The summed emission from these quasars generates the cosmic X-ray background, the spectrum of which has been used to argue that most black-hole growth is obscured^{1,2}. There are clear examples of obscured black-hole growth in the form of 'type-2' quasars^{3–5}, but their numbers are fewer than expected from modelling of the X-ray background. Here we report the direct detection of a population of distant type-2 quasars, which is at least comparable in size to the well-known unobscured type-1 population. We selected objects that have mid-infrared and radio emissions characteristic of quasars, but which are faint at near-infrared and optical wavelengths. We conclude that this population is responsible for most of the black-hole growth in the young Universe and that, throughout cosmic history, black-hole growth occurs in the dusty, gas-rich centres of active galaxies.

The population of unobscured type-1 quasars is now well understood out to redshift $z \approx 2$ by virtue of optical surveys⁶, which provide direct measurements around the 'break' in the luminosity function where the quasar population outputs most of its luminosity density. Our current understanding of the obscured type-2 quasar population is far less complete.

Examples of type-2 quasars have, for many years, been confined mainly to radio-selected samples. Radio emission is unaffected by dust and, at low frequencies, is not strongly dependent on viewing angle, so the simplest interpretation of optical spectroscopic follow-up of radio-selected samples is that luminous quasars divide roughly equally into obscured (type-2) and unobscured (type-1) objects⁷. However, 'radio-loud' quasars are rare and may be atypical: their large-scale radio jets may well modify their environments in ways which influence whether or not the nucleus is obscured, and hence the ratio of quasars in the type-1 and type-2 classes⁸.

More recently, deep X-ray surveys and optical surveys have also been successful in identifying type-2 quasars^{5,9,10}. However, optical spectroscopic follow-up has yet to find a type-2 population which can account fully for the X-ray background^{1,9,11}. This is believed to be because a substantial fraction of type-2 quasar nuclei are hidden by 'Compton-thick' material (with neutral gas-column densities $N_{\text{H}} \geq 10^{28} \text{ m}^{-2}$) even when observed at high X-ray energies^{1,12}.

Type-2 quasars (with visual extinction $A_V \geq 5$ magnitudes towards their nuclei) fail to outshine their host galaxies at ultraviolet and optical wavelengths. However, at longer wavelengths the obscuration becomes small, so the mid-infrared to radio properties of type-2 quasars should be similar to those of their type-1 counterparts, unless the obscuration or redshift becomes extreme. Indeed, a survey at 60 μm with the Infrared Astronomy Satellite (IRAS) led to

the discovery of the first high-redshift radio-quiet type-2 quasar: the $z = 2.3$ galaxy IRAS F10214+4724 (ref. 4). However, this object was discovered only because its flux was magnified by a factor of ~ 50 by a gravitational lens¹³. From the presence of high-excitation narrow lines and scattered broad lines¹⁴, it clearly has an obscured type-2 nucleus but its X-ray emission seems to be obscured by Compton-thick material¹².

Since the advent of the Spitzer Space Telescope¹⁵, with its sensitivity ≥ 100 times greater than IRAS and ≥ 10 times greater than the Infrared Space Observatory (ISO), type-2 quasars similar to IRAS F10214+4724 can be found from their mid-infrared emission without the aid of gravitational lenses. We chose the following selection criteria to find high-redshift type-2 quasars in the Spitzer First Look Survey (FLS): mid-infrared 24- μm flux density $S_{24\mu\text{m}} > 300 \mu\text{Jy}$; near-infrared 3.6- μm flux density $S_{3.6\mu\text{m}} \leq 45 \mu\text{Jy}$; and 350 $\mu\text{Jy} \leq S_{1.4\text{GHz}} \leq 2 \text{ mJy}$, where $S_{1.4\text{GHz}}$ is the 1.4-GHz radio flux density.

The mid-infrared ($S_{24\mu\text{m}}$) criterion was chosen to obtain a reliable (7- σ) catalogue from the Spitzer FLS data from the MIPS instrument (D.F. *et al.*, manuscript in preparation). At $z = 2$, this means the quasar luminosity will be $\geq 0.2L_{\text{quasar}}^*$ (where L_{quasar}^* is the break luminosity that corresponds to $M_B = -25.7$ magnitudes¹⁶); this means we can detect sources around the break in the luminosity function (see Supplementary Information). The observed flux at 24 μm corresponds to emitted flux at 8 μm at $z = 2$, and dust extinction is negligible at this wavelength unless the obscuring column is extreme. (We adopt a ΛCDM cosmology with the following parameters: $h = H_0/(100 \text{ km s}^{-1} \text{ Mpc}^{-1}) = 0.7$; $\Omega_m = 0.3$; $\Omega_\Lambda = 0.7$.)

The upper limit on the 3.6- μm flux (obtained using the IRAC instrument¹⁷) was chosen to select only the most distant type-2 quasars. At $z \geq 2$ the detected 3.6- μm flux density corresponds to light emitted at $\lambda \leq 1.2 \mu\text{m}$, so we are able to eliminate unobscured, type-1 quasars whose flux would exceed this limit (Fig. 1). Indeed, for type-2 quasars the $S_{3.6\mu\text{m}}$ emission is likely to be dominated by starlight from the host galaxy, allowing us to estimate 'photometric redshifts' z_{phot} (see Supplementary Information). Our $S_{3.6\mu\text{m}}$ criterion corresponds to $z_{\text{phot}} \geq 1.4$ (see footnote to Table 1 for more details).

The $S_{3.6\mu\text{m}}$ criterion includes objects without IRAC detections that we expect will have the highest redshifts. Because the 24- μm positions are accurate only to ~ 1 arcsec, the FLS radio positions¹⁸, accurate to ~ 0.5 arcsec, were better for spectroscopic follow-up. However, the main importance of the radio selection criteria is to ensure that the candidates are quasars rather than starburst galaxies. We chose a lower limit on $S_{1.4\text{GHz}}$ that is well above the level reached by high-redshift (submillimetre-selected) starburst galaxies without the benefit of gravitational lensing¹⁹, and an upper limit to filter out

¹Astrophysics, Department of Physics, University of Oxford, Keble Road, Oxford OX1 3RH, UK. ²Spitzer Science Center, California Institute of Technology, MS220-6, 1200 E. California Boulevard, Pasadena, California 91125, USA. ³Department of Physics, University of Durham, South Road, Durham DH1 3LE, UK. ⁴Herzberg Institute of Astrophysics, National Research Council, 5071 West Saanich Rd, Victoria, British Columbia V9E 2E7, Canada.

the radio-loud objects, whose extended jets might complicate interpretation.

Figure 1 shows the location of the 21 type-2 quasar candidates that met our selection criteria in the $S_{24\mu\text{m}}$ versus $S_{3.6\mu\text{m}}$ plane, and Table 1 summarizes their properties. The candidates were found from a sky area of 3.8 degrees² in the Spitzer FLS. No type-1 quasar ($A_V = 0$) that is bright enough to be detected at $24\mu\text{m}$ will be faint enough at $3.6\mu\text{m}$ to be selected (Fig. 1). ‘Blind’ low-resolution optical spectroscopy²⁰ was performed on our 21 candidates, ten of which yielded clear type-2 spectra with narrow emission lines giving redshifts in the range $1.4 \leq z \leq 4.2$. These objects are clearly type-2 quasars either because they have high-excitation lines (see the inset to Fig. 1) or because the rest-frame equivalent widths of their Ly- α lines are $>100\text{ nm}$ and are hence significantly larger than those seen in starbursts. Of the 11 objects that yielded no redshifts, all but one (which shows a faint red continuum) have completely blank spectra, showing that there is probably no contamination from lower-redshift ($z \leq 1.4$) starbursts, because they would have shown [O II] line emission, or if highly obscured, at least some continuum. We believe these blank-spectrum objects are also type-2 quasars with $z \geq 1.4$, and there is no compelling argument against them having $z \geq 2$ because the Ly- α line can easily be extinguished if the host galaxy is dusty on large (kiloparsec) scales.

We are probably seeing two types of type-2 quasar: the objects with narrow emission lines are obscured by the torus³, while the blank-spectrum objects are probably obscured by a starbursting host galaxy and we are seeing the coeval growth of a supermassive black hole and its host galaxy. Growing supermassive black holes embedded in a starburst have been found at $z \approx 2$ from X-ray measurements of submillimetre-selected galaxies²¹.

To interpret our results in terms of the ‘quasar fraction’ q —the ratio of the number of type-1 quasars to the total number of type-1 and type-2 quasars—we need to predict the average number $\langle N_1 \rangle$ of type-1 quasars meeting identical $S_{24\mu\text{m}}$ and $S_{1.4\text{GHz}}$ selection criteria, and having matched redshift and sky-area selection functions. We estimated a probability distribution $p(\langle N_1 \rangle)$ to account for the various uncertainties, the gaussian approximation of which ($4.3^{+2.2}_{-1.1}$) represents the expected number of type-1 quasars expected to adhere to both our $24\mu\text{m}$ and 1.4-GHz selection criteria²², and with $z \geq 2$ in a 3.8 degree² patch; this number would be 15 times higher without any radio-selection criteria.

Figure 2 shows the resulting conditional probability distribution of the quasar fraction q given our new data on type-2 quasars and our background knowledge of the type-1 quasar population, namely $p(q|\text{data}, \{\text{type-1 quasar}\})$. This bayesian approach allows us to account properly for small-number statistics (see Fig. 2 and Supplementary Information). It is difficult to judge whether the blank-spectrum objects with photometric redshifts $z_{\text{phot}} > 2$ should be included (dashed red line in Fig. 2) until we know more about the dust content of their host galaxies, but considering only the objects with spectroscopic redshifts (solid blue line in Fig. 2) is clearly a conservative estimate.

The most likely systematic problem with our analysis is that we have implicitly assumed identical radio properties²² for type-1 and type-2 quasars matched in intrinsic luminosity. However, ‘unified models’³ predict Doppler-boosted radio emission from the weak jets aligned closer to the line of sight for type-1 quasars²³ and hence, on average, lower radio luminosities for type-2 quasars. This would yield lower derived values of q and strengthen our conclusion that most quasars are obscured. It is of course possible to postulate that

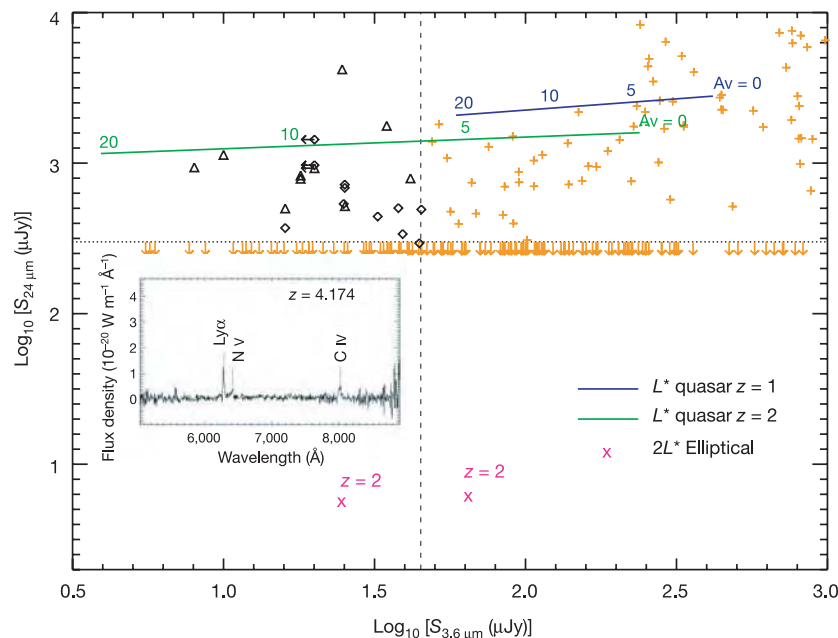


Figure 1 | Diagrammatic representation of the infrared selection criteria. $\text{Log}_{10}[S_{24\mu\text{m}} (\mu\text{Jy})]$ is plotted against $\text{Log}_{10}[S_{3.6\mu\text{m}} (\mu\text{Jy})]$. The dotted lines represent the flux density cuts of our selection criteria. Note that all objects plotted had to meet the criteria on radio flux density described in the text, and that the candidates were originally selected from the preliminary 3.6- and $24\mu\text{m}$ catalogues, but have had their fluxes slightly revised. This explains why two candidates are slightly outside the selection boundaries (see Supplementary Information). The parent population is plotted as orange crosses, or arrows when $S_{24\mu\text{m}}$ is an upper limit. Our 21 candidates are plotted in the top left corner, with black triangles for candidates with a spectroscopic redshift and black diamonds for candidates without one. Horizontal arrows represent upper limits for $S_{3.6\mu\text{m}}$. None of the candidates

showed type-1 quasar spectra. The spectrum of the $z = 4.174$ object is shown in the inset. The magenta crosses show $S_{24\mu\text{m}}$ and $S_{3.6\mu\text{m}}$ for a $2L^*$ elliptical galaxy at $z = 1$ and $z = 2$; starlight is assumed to evolve passively following a stellar population synthesis model³⁰. The other colours represent an L^* quasar (spectral energy distribution from ref. 31) with varying amounts of visual extinction A_V applied to $S_{24\mu\text{m}}$ and $S_{3.6\mu\text{m}}$ and plotted on the curves (assuming a Milky-Way-like extinction curve from ref. 32); these are plotted at two different redshifts, and the quasar population is assumed to undergo pure luminosity evolution¹⁶. At all redshifts, any type-1 ($A_V = 0$) quasar bright enough to make the $S_{24\mu\text{m}} > 300\mu\text{Jy}$ criterion will also be too bright at $3.6\mu\text{m}$ to make it into our sample. Inset, the optical spectrum of AMS16, showing Lyman- α , N V and C IV emission lines.

Table 1 | Basic data on the 21 type-2 quasars in our sample

Name	RA (J2000)	Dec.	$S_{24\mu\text{m}}$ (μJy)	$S_{1.4\text{GHz}}$ (μJy)	$S_{3.6\mu\text{m}}$ (μJy)	$S_{4.5\mu\text{m}}$ (μJy)	z_{phot}	z_{spec}	Spectroscopy
AMS01	17 13 11.17	+59 55 51.5	536	490	25.0	30.0	2.1		Blank
AMS02	17 13 15.88	+60 02 34.2	294	1184	44.5	61.0	1.4		Blank
AMS03	17 13 40.19	+59 27 45.8	500	1986	16.0	28.0	3.1	2.689	Single line*
AMS04	17 13 40.62	+59 49 17.1	828	536	18.0	22.0	2.8	1.782	
AMS05	17 13 42.77	+59 39 20.2	1769	1038	34.7	61.4	1.7	2.022	Weak Lyman- α
AMS06	17 13 43.91	+59 57 14.6	969	444	<20	<25	≥ 2.5		Blank
AMS07	17 14 02.25	+59 48 28.8	503	354	37.9	47.8	1.5		Blank
AMS08	17 14 29.67	+59 32 33.5	792	655	41.6	46.3	1.4	1.979	
AMS09	17 14 34.87	+58 56 46.4	685	426	25.2	<25	2.1		Blank
AMS10	17 16 20.08	+59 40 26.5	338	1645	39.2	44.7	1.5		Blank
AMS11	17 18 21.33	+59 40 27.1	442	356	32.4	55.6	1.7		Blank
AMS12	17 18 22.65	+59 01 54.3	518	946	25.24	<25	2.0	2.770	
AMS13	17 18 44.40	+59 20 00.8	4196	1888	24.7	49.1	2.1	1.986	
AMS14	17 18 45.47	+58 51 22.5	937	469	8.0	15.0	4.6	1.504	
AMS15	17 18 56.93	+59 03 24.1	371	440	16.0	16.0	3.0		Blank
AMS16	17 19 42.07	+58 47 08.9	788	390	18.0	21.0	2.8	4.174	
AMS17	17 20 45.17	+58 52 21.3	1134	615	10.0	15.0	3.9	3.138	Single line*
AMS18	17 20 46.32	+60 02 29.6	925	390	<20	<25	≥ 2.5	1.418	
AMS19	17 20 48.00	+59 43 20.7	1433	822	<20	26.9	≥ 2.5		Blank
AMS20	17 20 59.10	+59 17 50.5	492	1268	45.2	64.9	1.4		Faint red continuum
AMS21	17 21 20.09	+59 03 48.6	720	449	25.2	38.6	2.1		Blank

The J2000.0 positions are from the Spitzer FLS radio catalogue¹⁸. RA, right ascension, Dec., declination. The MIPS 24- μm flux density $S_{24\mu\text{m}}$ is obtained by point-spread function (PSF) fitting²⁹ as all objects are point sources at the ~ 6 -arcsecond resolution of the MIPS observations, with positional errors of ~ 1 arcsec. It has a typical error of ± 10 –15%. The 1.4-GHz flux density is the peak value (in μJy per beam) from the radio catalogue¹⁸. The IRAC 3.6- and 4.5- μm flux densities are measured in 5-arcsec-diameter apertures¹⁷ and have typical errors of $\pm 10\%$. The photometric redshifts are calculated by assuming that the 3.6- μm flux density is dominated by starlight from a $2L_{\text{gal}}$ galaxy (L_{gal} is the break in the galaxy luminosity function, following the models of ref. 30 and passive evolution). The criterion $S_{3.6\mu\text{m}} \leq 45 \mu\text{Jy}$ corresponds to a limiting photometric redshift of $z_{\text{phot}} \approx 1.4$; this was chosen to allow for scatter in the photometric redshift estimation, while still filtering out type-1 quasars and low-redshift contaminants like radio galaxies. All of the 21 candidates were observed spectroscopically with the dual-beam ISIS instrument at the William Herschel Telescope, in July 2004 and April 2005. 'Blind' low-resolution spectroscopy²⁰ was performed by offsetting from nearby stars to the radio positions, with exposure times of ~ 30 min and a continuous wavelength coverage across the entire visible band. Objects with an asterisk have only one definite line in their optical spectra, which is taken to be Ly- α by virtue of extreme equivalent width, blue-absorbed line profile and a lack of any other lines supporting alternative identifications.

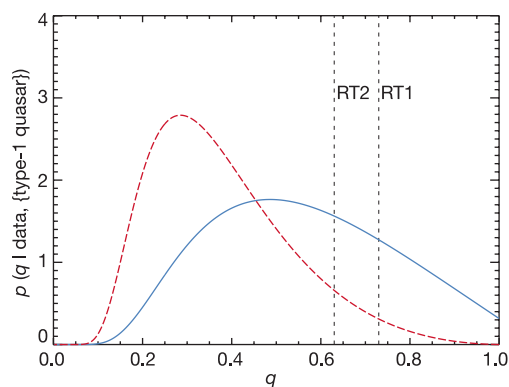


Figure 2 | Probability distributions for the quasar fraction. The normalized posterior probability $p(q|\text{data}, \{\text{type-1 quasar}\})$ of the quasar fraction q given our new data on type-2 quasars, and background information on type-1 quasars. The two lines represent different interpretations of the number of type-2 quasars at $z \geq 2$ determined from our new data. To obtain $p(q|\text{data}, \{\text{type-1 quasar}\})$ we used Bayes' theorem with a uniform prior over the range $0 \leq q \leq 1$ so that the $p(q|\text{data}, \{\text{type-1 quasar}\}) = p(\text{data}|q, \{\text{type-1 quasar}\})$. We then calculated this likelihood function, at each q , as a Poisson distribution for the observed (integer) number of type-2 quasars at $z \geq 2$ where the mean number expected $\langle N_2 \rangle = (1 - q) \langle N_1 \rangle / q$. Because of uncertainties in the background information on $\langle N_1 \rangle$, we had to evaluate the likelihood at each q and $\langle N_1 \rangle$ and then integrate $p(N_1)$ times the likelihood over $\langle N_1 \rangle$. The solid blue line is the posterior probability if we believe the only candidates with $z \geq 2$ are the five whose spectroscopic redshifts confirm this. If, additionally, we use the photometric redshifts of the candidates with blank spectra (six additional candidates with $z \geq 2$), then we obtain the red curve. The best estimate for q is probably somewhere between the two curves. The two vertical lines represent the predictions of receding torus models, one with a fixed torus height (marked 'RT1')²⁶ and one ('RT2') with a torus height varying with luminosity²⁷.

the optical-to-radio correlation evolves with redshift, but this is currently unknown (see also Supplementary Information).

The median redshift of our spectroscopic sample is at $z = 2$ which is exactly where the optical luminosity density of the type-1 quasar population is at its peak⁶. Our sample shows that during this 'epoch of quasar activity' we can be confident that about half of all high-luminosity quasars are obscured, and if some of the blank-spectrum objects turn out to have $z \geq 2$ then it is likely that type-2 quasars outnumber type-1 quasars by a significant factor. It is also possible that we are missing type-2 quasars with an extreme obscuring column ($A_V \approx 100$), which could be due, for example, to the obscuring torus being completely edge-on to the line of sight, but this would only strengthen our result, which has been predicted both by models for the X-ray background¹, and by models predicting the surface density of faint AGN in relatively small-sky-area surveys²⁴. This is, however, the first direct detection of a high-redshift type-2 quasar population that is likely to outnumber the type-1 population. We note that the solid blue line (which compares narrow-line or 'torus-obscured' type-2 quasars with type-1 quasars) is consistent with the quasar fraction from radio-loud samples⁷, while the dashed red line (which includes both torus-obscured and 'host-obscured' type-2 quasars) has enough obscured objects to account for the X-ray background¹.

Figure 2 shows the expected ratio for two 'receding torus'^{25,26,27} models in which the solid angle subtended by the obscuring torus decreases (so that q increases) systematically with increasing quasar luminosity. Because we have suggested that some of the blank-spectrum objects are observed by kiloparsec-scale dust, it is only the ratio of spectroscopically confirmed type-2 quasars to type-1 quasars (the solid blue line) that is relevant to the model predictions, and so these models are not excluded.

To conclude, we consider our results alongside those of the X-ray surveys at lower redshifts⁹ ($z \leq 2$), which find that most accretion at low redshift occurs in type-2 quasars and type-2 Seyferts (their lower-luminosity analogues). We deduce that throughout cosmic history, black-hole growth has been concentrated in obscured regions. This is in good agreement with predictions from the X-ray background¹ and

implies, from comparisons between the integrated luminosity density of quasars (both type-1 and type-2) and the local space density of relic black holes²⁸, that black-hole growth occurs in short, efficient spurts in the cores of forming galaxies².

Received 24 December 2004; accepted 18 May 2005.

1. Worsley, M. A., Fabian, A. C., Barcons, X., Mateos, S. & Hasinger, G. The (un)resolved X-ray background in the Lockman Hole. *Mon. Not. R. Astron. Soc.* **354**, 720–726 (2004).
2. Fabian, A. C. The obscured growth of massive black holes. *Mon. Not. R. Astron. Soc.* **308**, L39–L43 (1999).
3. Antonucci, R. Unified models for active galactic nuclei and quasars. *Annu. Rev. Astron. Astrophys.* **31**, 473–521 (1993).
4. Rowan-Robinson, M. *et al.* A high-redshift IRAS galaxy with huge luminosity—hidden quasar or protogalaxy? *Nature* **351**, 719–721 (1991).
5. Norman, C. *et al.* A classic Type 2 quasar. *Astrophys. J.* **571**, 218–225 (2002).
6. Wolf, C. *et al.* The evolution of faint AGN between $z \sim 1$ and $z \sim 5$ from the COMBO-17 survey. *Astron. Astrophys.* **408**, 499–514 (2003).
7. Willott, C. J., Rawlings, S., Blundell, K. M. & Lacy, M. The quasar fraction in low-frequency-selected complete samples and implications for unified schemes. *Mon. Not. R. Astron. Soc.* **316**, 449–458 (2000).
8. Baker, J. C. *et al.* Associated absorption in radio quasars. I. C IV absorption and the growth of radio sources. *Astrophys. J.* **568**, 592–609 (2002).
9. Barger, A. J. *et al.* The cosmic evolution of hard X-ray-selected active galactic nuclei. *Astron. J.* **129**, 578–609 (2005).
10. Zakamska, N. L. *et al.* Candidate type II quasars from the SDSS: III. Spectropolarimetry reveals hidden type I nuclei. *Astron. J.* (in the press).
11. Zheng, W. *et al.* Photometric redshifts of X-ray sources in the Chandra Deep Field South. *Astrophys. J. Suppl. Ser.* **155**, 73–87 (2004).
12. Alexander, D. M. *et al.* A Chandra observation of the $z = 2.285$ galaxy FSC10214+4724: Evidence for a Compton-thick quasar? *Mon. Not. R. Astron. Soc.* **357**, L16–L20 (2005).
13. Broadhurst, T. & Lehar, J. A gravitational lens solution for the IRAS Galaxy FSC 10214+4724. *Astrophys. J.* **450**, L41–L44 (1995).
14. Serjeant, S. *et al.* A spectroscopic study of IRAS F10214+4724. *Mon. Not. R. Astron. Soc.* **298**, 321–331 (1998).
15. Werner, M. W. *et al.* The Spitzer Space Telescope Mission. *Astrophys. J. Suppl. Ser.* **154**, 1–9 (2004).
16. Croom, S. M. *et al.* The 2dF QSO Redshift Survey-XII. The spectroscopic catalogue and luminosity function. *Mon. Not. R. Astron. Soc.* **349**, 1397–1418 (2004).
17. Lacy, M. *et al.* The Infrared Array Camera component of the Spitzer Space Telescope Extragalactic First Look Survey. *Astrophys. J. Suppl. Ser.* (in the press).
18. Condon, J. J. *et al.* The SIRTIF First-Look survey. I. VLA image and source catalog. *Astrophys. J.* **125**, 2411–2426 (2003).
19. Chapman, S. C., Blain, A. W., Smail, I. & Ivison, R. J. A redshift survey of the submillimetre galaxy population. *Astrophys. J.* **622**, 772–796 (2005).
20. Rawlings, S., Eales, S. & Warren, S. The detection of four high-redshift ($0.5 \leq z \leq 3.22$) radiogalaxies by optical spectroscopy of five blank fields. *Mon. Not. R. Astron. Soc.* **243**, 14–18 (1990).
21. Alexander, D. M. *et al.* Rapid growth of black holes in massive star-forming galaxies. *Nature* **434**, 738–740 (2005).
22. Cirasuolo, M., Celotti, A., Magliocchetti, M. & Danese, L. Is there a dichotomy in the radio loudness distribution of quasars? *Mon. Not. R. Astron. Soc.* **346**, 447–455 (2003).
23. Miller, P., Rawlings, S. & Saunders, R. The radio and optical properties of the $z < 0.5$ BQS quasars. *Mon. Not. R. Astron. Soc.* **263**, 425–460 (1993).
24. Treister, E. *et al.* Obscured AGN and the X-ray, optical and far-infrared number counts of AGN in the GOODS fields. *Astrophys. J.* **616**, 123–135 (2004).
25. Lawrence, A. The relative frequency of broad-lined and narrow-lined active galactic nuclei—implications for unified schemes. *Mon. Not. R. Astron. Soc.* **252**, 586–592 (1991).
26. Simpson, C. A new look at the isotropy of narrow-line emission in extragalactic radio sources. *Mon. Not. R. Astron. Soc.* **297**, L39–L43 (1998).
27. Simpson, C. The luminosity dependence of the type 1 active galactic nucleus fraction. *Mon. Not. R. Astron. Soc.* **360**, 565–572 (2005).
28. Yu, Q. & Tremaine, S. Observational constraints on growth of massive black holes. *Mon. Not. R. Astron. Soc.*, 965–976 (3352002).
29. Marleau, F. R. *et al.* Extragalactic source counts at 24 microns in the Spitzer First Look survey. *Astrophys. J. Suppl. Ser.* **154**, 66–69 (2004).
30. Bruzual, G. & Charlot, S. Stellar population synthesis at the resolution of 2003. *Mon. Not. R. Astron. Soc.* **344**, 1028 (2003).
31. Rowan-Robinson, M. A new model for the infrared emission of quasars. *Mon. Not. R. Astron. Soc.* **272**, 737–748 (1995).
32. Pei, Y. C. Interstellar dust from the Milky Way to the Magellanic Clouds. *Astrophys. J.* **395**, 130–139 (1992).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank C. Wolf, L. Clewley, H.-R. Klöckner and G. Cotter for discussions. A.M.-S. is grateful to the Council of the European Union for financial support. S.R. and C.S. are grateful to the UK PPARC for a Senior Research Fellowship and an Advanced Fellowship respectively. Q7

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to A.M.-S. (a.martinez-sansigre1@physics.oxford.ac.uk).

LETTERS

No oceans on Titan from the absence of a near-infrared specular reflection

R. A. West¹, M. E. Brown², S. V. Salinas³, A. H. Bouchez⁴ & H. G. Roe²

With its substantial atmosphere of nitrogen, hydrocarbons and nitriles, Saturn's moon Titan is a unique planetary satellite. Photochemical processing of the gaseous constituents produces an extended haze that obscures the surface. Soon after the Voyager fly-bys in 1980 and 1981 photochemical models^{1–3} led to the conclusion that there should be enough liquid methane/ethane/nitrogen to cover the surface to a depth of several hundred metres. Recent Earth-based radar echoes imply that surface liquid may be present at a significant fraction of the locations sampled⁴. Here we present ground-based observations (at near-infrared wavelengths) and calculations showing that there is no evidence thus far for surface liquid⁵. Combined with the specular signatures from radar observations, we infer mechanisms that produce very flat solid surfaces, involving a substance that was liquid in the past but is not in liquid form at the locations we studied.

Specular reflection from a surface that is smooth on the scale of the wavelength of the observation can be used to diagnose surface liquid. We searched for specular reflection in high-resolution images of Titan obtained on 20 nights between 17 September 2003 and 11 January 2004, and 11 nights between 2 September 2004 and 3 November 2004. On the four nights of 24–27 December 2004 we observed Titan continuously for approximately ten hours each night, allowing a much more precise search for changing specular reflection. We scaled the intensity of each image relative to the others by scaling to the brightness of the outer edge of the northern limb of Titan, which is only affected by the relatively unchanging haze and not by clouds or surface features.

No specular reflection is obvious in visual inspection of any of the images. The locations sampled in these images are shown in Fig. 1. To search for fainter specular reflection in the extensive December 2003 images, we projected each image onto a cylindrical latitude–longitude map to examine the change in surface brightness of individual locations on the surface of Titan with time as Titan rotated. These differential phase curves for the four regions that could have shown specular reflection are shown in Fig. 2.

To understand how sensitive the near-infrared images are for the detection of a specular signature, we simulated the radiative transfer of photons scattered from Titan's haze and surface. To produce these calculations we modified a public-domain code⁶ to compute the intensity. The code includes multiple scattering from a plane-parallel multilayer haze and scattering from a surface with a Lambert (isotropic) reflection model. Our modification follows a formalism for the optical properties of a distribution of wave slopes and facet orientations used by the terrestrial sea-surface community⁷. Multiple surface–atmosphere reflections are included just as they are for a Lambert surface. Accuracy to better than 1% was verified by comparison with a benchmark calculation (H. Gordon, personal communication, 1994).

Campbell *et al.*⁴ report a specular signature at 12 of the 16 locations they sampled with the Arecibo radar. For these they derived root-mean-square (r.m.s.) slope values between 0.5 and 3.5 degrees with a median value of ~2 degrees and uncertainty typically 20% of the mean value. In this work we calculated the specular signature for r.m.s. slope values of 3.1, 4.9 and 5.5 degrees. The results are shown in Fig. 3 for these three r.m.s. slope values and for two wavelengths (2.1 and 0.938 μm) with appropriate differences in haze optical depth and scattering asymmetry parameter. In terms of I/F , where I is the reflected intensity and πF is the incident solar flux, the specular signature (the peak at the centre of the disk) has a full-width at half-maximum (FWHM) value of about 5.5% of Titan's diameter and a peak value of $I/F = 0.37$ for the case with optical depth $\tau = 0.3$ and r.m.s. slope 3.1 degrees. The FWHM is 11% of Titan's diameter and the maximum I/F is 0.15 for the higher r.m.s. slope value of 5.5 degrees. The signatures are weaker for the higher optical depth case

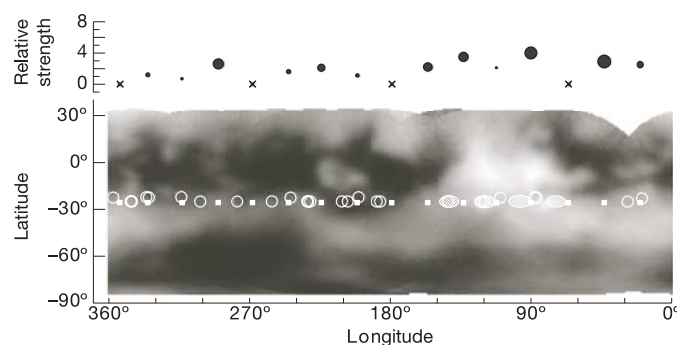


Figure 1 | A cylindrical projection map of Titan at 2- μm wavelength¹¹. The bright Xanadu region appears near the equator at 90 degrees longitude. The small solid squares show the location of the radar observations of ref. 4. The open circles show the region of expected specular reflection for our 27 Keck images. All observations were obtained with the Keck II adaptive optics system and NIRC2 near-infrared camera and spectrograph¹², with an angular resolution of approximately 0.049 arcsec at 2.1 μm , corresponding to a spatial resolution on Titan of between 300 and 400 km. On most nights we obtained a single set of three-position dithered 30-s exposures in the K' (1.95–2.30 μm) filter. No more than 15% of the area of the open circles could be covered with material with an r.m.s. smoothness comparable to that measured by radar. The hatched regions near 142, 119, 96 and 74 degrees longitude show the region covered by intensive Keck observations where the upper limits to specular reflection are even stronger. The upper panel shows the relative strength of the radar specular reflection, while the size of the symbol used shows approximately the size of the region from which the specular signature came and is proportional to the inferred r.m.s. slope. An 'x' means that no radar specular reflection was visible at that location.

¹MS 169-237 Jet Propulsion Laboratory, 4800 Oak Grove Drive, Pasadena, California 91109, USA. ²Division of Geological and Planetary Sciences, California Institute of Technology, Pasadena, California 91125, USA. ³Centre for Remote Imaging, Sensing and Processing, National University of Singapore, Blk SOC 1, Level 2, Lower Kent Ridge Rd, Singapore 119260. ⁴W. M. Keck Observatory, 65-1120 Mamalahoa Highway, Kamuela, Hawaii 96743, USA.

but the FWHM values are unchanged. The values quoted above apply to an observation (by the Cassini Visible and Infrared Mapping Spectrometer⁸, or VIMS) capable of fully resolving these features. The point-spread function of our ground-based images reduces the peak as shown in Fig. 3.

A specular reflection as strong as those calculated for the optical thickness at $2.1\text{ }\mu\text{m}$ would immediately be obvious in any of the Keck images. A surface that is composed of only a fractional covering of a smooth surface would have a smaller specular reflection, however. To place limits on the smooth surface fraction we added simulated specular reflections to each of the images and determined when the specular reflection was obvious in the image. For the median r.m.s. surface slopes found by Arecibo of ~ 2 degrees, we can firmly rule out smooth (at $2.1\text{ }\mu\text{m}$) surfaces covering more than 15% of any of the 31 regions surveyed. A similar assessment for high-resolution Cassini VIMS images yields a limit of 2% for the same r.m.s. slope.

A stronger limit can be placed on the specular locations for the four nights of extensive observation. We again add simulated specular reflections to each image from those nights, but now quantitatively examine the brightness as a function of solar phase angle, with and without the specular reflection. For the r.m.s. surface slope 3.1 -degree model, we place upper limits to the smooth fraction of the potential specular surface of 3, 8, 10 and 8% for the regions at 142 , 119 , 96 and 74 degrees longitude, respectively. A similar analysis of high-resolution Cassini VIMS images should provide tighter limits by about a factor of four.

The upper limits we place on the fraction of Titan's surface covered by liquid challenge the liquid-surface interpretation suggested by the Arecibo observations. It is possible that the radar observations sampled lakes of liquid hydrocarbons as small as several tens of kilometres in diameter that were not seen in the infrared images. Our results suggest that at most a small percentage of the surface of Titan studied thus far could be covered by lakes. Alternatively, we must ask what processes might form a surface that is sufficiently flat and smooth at a wavelength of 13 cm to produce the specular signatures seen in the Arecibo echoes, and which is at the same time rough enough to inhibit a specular signature at $2.1\text{ }\mu\text{m}$. We propose three possible mechanisms: evaporite deposits, ponding of an inviscid fluid which then froze, and surface accumulation of sedimented haze from the stratosphere. All of these involve liquid on the surface at

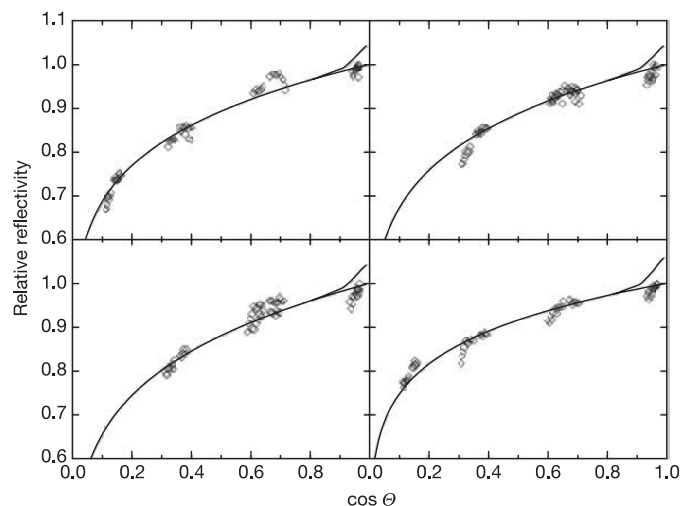


Figure 2 | The $2.1\text{-}\mu\text{m}$ brightness of four locations on the surface of Titan measured over four days as a function of solar zenith angle. The locations are at longitudes 142 , 119 , 96 and 74 degrees and are shown as hatched regions in Fig. 1. The relative intensities are fitted to a function of the form $\cos \theta$, where θ is the solar zenith angle. The small deviations from these lines near $\cos \theta = 1$ show the expected effect of adding a specular reflection at the stated upper limit and represent our detection limits.

some time in the past but not the present.

A fourth possibility sometimes considered—that a thin surface layer of solid particles is floating on a liquid—is problematic because (1) the density of a solid hydrocarbon or water ice particle is likely to be higher than that of liquid methane/ethane/nitrogen and (2) the surface tension of liquid methane is very low. The first of these difficulties could be overcome if the solid hydrocarbon were structured as a fluffy aggregate, which indeed the sedimenting particles are thought to be⁹.

Evaporite deposits on the Earth are a class of sedimentary minerals and sedimentary rocks that form by precipitation from evaporating

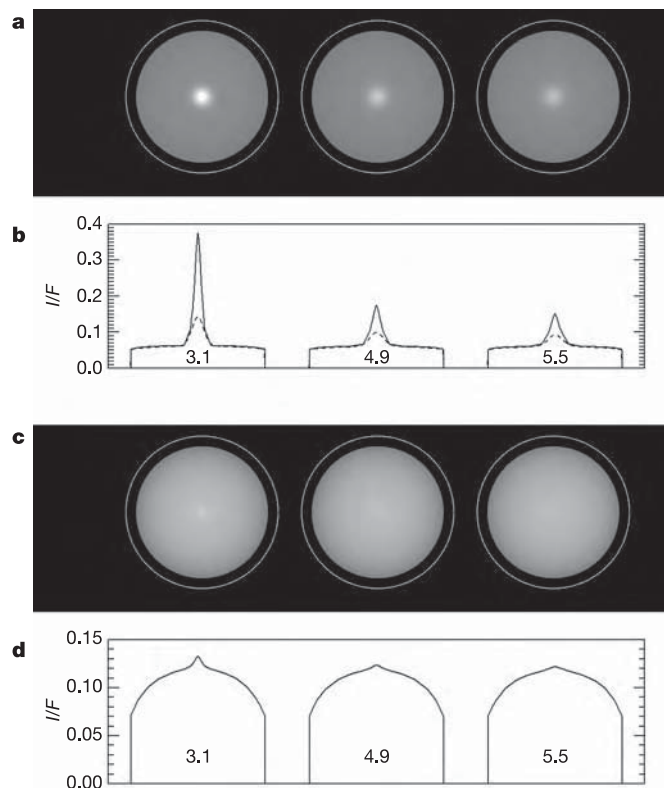


Figure 3 | Synthesized images and plots of reflectivity. The perspective is for Titan at opposition. The plots cut through the centre of each image. Panels **a** and **b** apply to the $2.1\text{-}\mu\text{m}$ window where the optical depth is close to 0.3 and the haze scattering asymmetry parameter is 0.2 . The dashed curve was produced by convolving the synthetic image with the Keck point-spread function. The solid curves apply to Cassini $2.1\text{-}\mu\text{m}$ images near Titan where the spatial resolution is much better than 300 km . Panels **c** and **d** are for the Cassini Imaging Science Subsystem images in the $0.938\text{-}\mu\text{m}$ window where the optical depth is 3 and the haze asymmetry parameter is close to 0.75 . Grey levels are calculated for view angles not more than 60 degrees to avoid errors near the limb from the plane-parallel approximation. The white circles surrounding the image show the location of the solid surface radius. The labels under the specular feature indicate the r.m.s. slope. Atmospheric scattering properties are determined by the nature of the haze particles. The optical depth of the haze may be as high as 3 at $0.938\text{ }\mu\text{m}$, and close to 0.3 near $2.1\text{ }\mu\text{m}$ (ref. 13). We computed particle scattering phase functions and single-scattering albedos at $0.938\text{ }\mu\text{m}$, based on a fractal aggregate model^{14,15} with optical constants of Titan 'tholin'¹⁶ at $0.938\text{ }\mu\text{m}$ but with the imaginary part multiplied by $4/3$ to match the current geometric albedo measurements better. The derived optical depth at $2.1\text{ }\mu\text{m}$ (ref. 13) was also based on a fractal aggregate model with scattering phase function asymmetry (cosine-weighted angular distribution) $g = 0.2$. Results for $2.1\text{ }\mu\text{m}$ reported here used a Henyey-Greenstein phase function ($P(g, \theta) = (1 - g^2)/(1 + g^2 - 2g \cos \theta)^{3/2}$) with $g = 0.2$. The Fresnel reflection coefficients are derived from the complex refractive index of the liquid. We used a value of $(1.286, 0.114)$ as representative for a mixture of liquid methane/ethane near 90 K (ref. 17).

fluid. Common terrestrial evaporite minerals are halite, gypsum and anhydrite, which can form as sea water evaporates, and the rocks limestone and dolostone. On Titan the composition would probably be a hydrocarbon layer from fully or partially dissolved and undissolved organic material that entered a liquid methane/ethane/nitrogen body from an accumulation of sedimenting organic haze or geothermal sources. If the evaporite material were porous it could contain a significant amount of subsurface liquid that would be available to replenish methane lost to photo-dissociation.

A second possible mechanism involves volcanic extrusion of liquid water/ammonia which ponded to form a very flat surface and then froze. This is less likely to be the mechanism producing the Arecibo echoes, however, because the radar cross-sections are more consistent with organic material. This objection may be overcome by a surface deposit of organic sediment, a specific example of our next proposal.

A third possibility is that deposits of organic haze over geologic timescales have filled in topographically low areas with a flat top facilitated by liquid methane/ethane/nitrogen that later evaporated, or by wind action. These deposits may well be porous and contain significant amounts of subsurface liquid methane/ethane/nitrogen that now buffers the atmosphere.

Although this possibility and the other mechanisms we propose may succeed in accounting for a solid surface smooth enough to produce a specular signature at 13 cm, we point out that any solid surface will roughen at 13-cm scales over geologic time owing to a variety of geologic or aeolian (wind-related) processes, and so we may need to invoke a geologically young age as well as one of these mechanisms to explain the surfaces we see.

In addition, current large-scale cloud activity on Titan is rare outside the southern high latitudes but is predicted to move north as Titan's southern summer turns to southern autumn¹⁰, suggesting the possibility that these regions are seasonally refreshed with precipitation, which could work to smooth the surface. Unless there is a subsurface reservoir capable of sustaining atmospheric methane, the long-term prognosis is for a Titan atmosphere depleted in methane, having profound implications for photochemistry, radiative balance and circulation.

Received 10 February; accepted 18 May 2005.

1. Flasar, F. M. Oceans on Titan. *Science* **221**, 55–57 (1983).

2. Lunine, J. I., Stevenson, D. J. & Yung, Y. L. Ethane ocean on Titan. *Science* **222**, 1229–1230 (1983).
3. Yung, Y. L., Allen, M. & Pinto, J. P. Photochemistry of the atmosphere of Titan: Comparison between model and observations. *Astrophys. J. Suppl.* **55**, 465–506 (1984).
4. Campbell, D. B., Black, G. J., Carter, L. M. & Ostro, S. J. Radar evidence for liquid surfaces on Titan. *Science* **302**, 431–434 (2003).
5. Porco, C. C. *et al.* Imaging Titan from the Cassini Spacecraft. *Nature* **434**, 159–168 (2005).
6. Evans, K. F. & Stephens, G. L. A new polarized radiative transfer model. *J. Quant. Spectrosc. Radiat. Trans.* **46**, 413–423 (1991).
7. Deuze, J. L., Herman, M. & Santer, R. Fourier series expansion of the transfer equation in the atmosphere-ocean system. *J. Quant. Spectrosc. Radiat. Trans.* **41**, 483–494 (1989).
8. Brown, R. H. *et al.* The Cassini Visual and Infrared Mapping Spectrometer (VIMS) investigation. *Space Sci. Rev.* **115**, 111–168 (2004).
9. West, R. A. & Smith, P. H. Evidence for aggregate particles in the atmospheres of Titan and Jupiter. *Icarus* **90**, 330–333 (1991).
10. Brown, M. E., Bouchez, A. H. & Griffith, C. A. Direct detection of variable tropospheric clouds near Titan's south pole. *Nature* **420**, 795–797 (2002).
11. Bouchez, A. H., Brown, M. E., Schaller, E. L. & Roe, H. G. Cloud frequency and wind speed in Titan's troposphere. *Bull. Am. Astron. Soc.* **36**, 1696 (2004).
12. Wizinowich, P. *et al.* Performance of the W.M. Keck observatory natural guide star adaptive optic facility: the first year at the telescope. *Proc. SPIE* **4007**, 2–13 (2000).
13. Gibbard, S. G. *et al.* Titan's 2 μ m surface albedo and haze optical depth in 1996–2004. *Geophys. Res. Lett.* **32**, L17S02 (2004).
14. Lemmon, M. T. *Properties of Titan's Haze and Surface*. PhD thesis, Arizona Univ. (1994).
15. Grieger, B., Lemmon, M. T., Markiewicz, W. J. & Keller, H. U. Inverse radiation modeling of Titan's atmosphere to assimilate solar aureole imager data of the Huygens probe. *Planet. Space Sci.* **51**, 147–158 (2003).
16. Khare, B. N. *et al.* Optical constants of organic tholins produced in a simulated Titanian atmosphere—from soft-x-ray to microwave-frequencies. *Icarus* **60**, 127–137 (1984).
17. Martonchik, J. V. & Orton, G. S. Optical constants of liquid and solid methane. *Appl. Opt.* **33**, 8306–8317 (1994).

Acknowledgements We benefited from radiative transfer benchmark calculations provided by H. Gordon and from discussions with D. Campbell, R. Lorenz and S. Ostro. Part of this work was performed by the Jet Propulsion Laboratory, California Institute of Technology under contract with the National Aeronautics and Space Administration.

Author Contributions M.E.B., A.H.B. and H.G.R. supplied the Keck observations and data analysis. S.V.S. provided the light-scattering model calculations.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.A.W. (Robert.A.West@jpl.nasa.gov).

Partial quantum information

Michał Horodecki¹, Jonathan Oppenheim² & Andreas Winter³

Information—be it classical¹ or quantum²—is measured by the amount of communication needed to convey it. In the classical case, if the receiver has some prior information about the messages being conveyed, less communication is needed³. Here we explore the concept of prior quantum information: given an unknown quantum state distributed over two systems, we determine how much quantum communication is needed to transfer the full state to one system. This communication measures the partial information one system needs, conditioned on its prior information. We find that it is given by the conditional entropy—a quantity that was known previously, but lacked an operational meaning. In the classical case, partial information must always be positive, but we find that in the quantum world this physical quantity can be negative. If the partial information is positive, its sender needs to communicate this number of quantum bits to the receiver; if it is negative, then sender and receiver instead gain the corresponding potential for future quantum communication. We introduce a protocol that we term ‘quantum state merging’ which optimally transfers partial information. We show how it enables a systematic understanding of quantum network theory, and discuss several important applications including distributed compression, noiseless coding with side information, multiple access channels and assisted entanglement distillation.

In any given situation, we can ask: how much is there to know? And, how large is our ignorance? The formulation of these questions addresses the quantity of information, not its content, which is simply because the latter is hard to assess and to compare. The former approach to classical information was pioneered by Shannon¹, who provided the tools and concepts to answer the first of the two questions: the amount of information originating from a source is the memory required to faithfully represent its output. In the case discussed in this work—of a statistical source—this amount is given by its entropy.

To approach the second question, let us consider a two-player game. One participant (Bob) has some incomplete prior information Y , the other (Alice) holds some information X : we think of X and Y as random variables, and Bob has prior information owing to possible correlations between X and Y . If Bob wants to learn X , how much additional information does Alice need to send him? This is one of the key problems of classical information theory, because it describes a ubiquitous scenario in information networks. It was solved by Slepian and Wolf², who proved that the amount of information that Bob needs is given by a quantity called the ‘conditional entropy’. It measures the partial information that Alice must send to Bob so that he gains full knowledge of X given his previous knowledge from Y , and it is just the difference between the entropy of (X, Y) taken together (the total information) and the entropy of Y (the prior information). This result is remarkable because Alice communicates without knowing what Bob already knows: she only needs to know her distribution, and the value of the conditional entropy. Of course, this partial information is always a positive quantity. Classically, there would be no meaning to negative information.

In the quantum world, the first of the two questions, how to quantify quantum information, was answered by Schumacher², who showed that the minimum number of quantum bits required to convey quantum information is given by the quantum (von Neumann) entropy. To answer the second question, we now consider the quantum version of the two-party scenario above: Alice and Bob each possess a system in some unknown quantum state with the total density operator being ρ_{AB} and each party having states with density operators ρ_A and ρ_B respectively. In the case where Bob is correlated with Alice, he has some prior information about her state. We now ask how much additional quantum information Alice needs to send him, so that he has the full state (with density operator ρ_{AB}). Because we want to quantify the quantum partial information, we are interested in the minimum amount of quantum communication to do this, allowing unlimited classical communication—the latter type of information being far easier to transmit than the former, as it can be sent over a telephone, whereas the former is extremely delicate and must be sent using a special quantum channel.

We are interested in information theoretic (asymptotic) quantities, in the spirit of Shannon, so we go to the limit of many copies of state ρ_{AB} and vanishing but non-zero errors in the protocol. We find here that the amount of partial quantum information that Alice needs to send Bob is given by the quantum conditional entropy, which is exactly the same quantity as in the classical case but with the Shannon entropy changed to the von Neumann entropy:

$$S(A|B) \equiv S(AB) - S(B) \quad (1)$$

where $S(B)$ is the entropy of Bob’s state ρ_B and $S(AB)$ is the entropy of the joint state ρ_{AB} . For quantum states, the conditional entropy can be negative^{4–6}, and thus it is rather surprising that this quantity has a physical interpretation in terms of how much quantum communication is needed to gain complete quantum information: that is, possession of a system in the total state ρ_{AB} .

However, in the above scenario, the negative conditional entropy can be clearly interpreted. We find that when $S(A|B)$ is negative, Bob can obtain the full state using only classical communication, and additionally, Alice and Bob will have the potential to transfer additional quantum information in the future at no further cost to them. They end up sharing $-S(A|B)$ Einstein–Podolsky–Rosen (EPR) pairs⁷, that is, pure maximally entangled states $\frac{1}{\sqrt{2}}(|00\rangle + |11\rangle)$, which can be used to teleport⁸ quantum states between the two parties using only classical communication. Negative partial information thus also gives Bob the potential to receive future quantum information for free. The conditional entropy plays the same role in quantum information theory as it does in the classical theory, except that here, the quantum conditional entropy can be negative in an operationally meaningful way. We can say that the ignorance of Bob—the conditional entropy—if it is negative, precisely cancels the amount by which he knows too much⁹; the negative conditional entropy is just the potential future communication gained.

This solves the well-known puzzle of how to interpret the quantum conditional entropy, which has persisted despite interesting attempts

¹Institute of Theoretical Physics and Astrophysics, University of Gdańsk, 80-952 Gdańsk, Poland. ²Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge CB3 0WA, UK. ³Department of Mathematics, University of Bristol, Bristol BS8 1TW, UK.

to understand it⁶. Because there are no conditional probabilities for quantum states, $S(A|B)$ is not an entropy as in the classical case. But by going back to the definition of information in terms of storage space needed to hold a message or state, we can make operational sense of this quantity.

We now turn to the protocol which allows Alice to transfer her state to Bob's site in the above scenario (we henceforth adopt the common usage of referring directly to manipulations on a 'state'—meaning a manipulation on a physical system in some quantum state). We call this transfer 'quantum state merging', because Alice is effectively merging her state with that of Bob's. We note that in quantum information theory, faithful state transmission means that while the state merging protocol may depend on the density operator of the source, it must succeed with high probability for any pure state sent. An equivalent and elegant way of expressing this criterion is to imagine that ρ_{AB} is part of a pure state $|\psi\rangle_{ABR}$, which includes a reference system R . Alice's goal is to transfer the state ρ_A to Bob, and we demand that after the protocol, the total state still has high fidelity with $|\psi\rangle_{ABR}$ (meaning they are nearly identical); see Fig. 1 which includes a high-level description of the protocol. The essential element of state merging is that ρ_R must be unchanged, and Alice must decouple her state from R . This also means (seemingly paradoxically) that as far as any outside party is concerned, neither the classical nor quantum communication is coupled with the merged state.

Let us consider three simple and instructive examples:

- (1) Alice has a completely unknown state which we can represent as the maximally mixed density matrix $\rho_A = \frac{1}{2}(|0\rangle\langle 0|_A + |1\rangle\langle 1|_A)$, and Bob has no state (or a known state $|0\rangle_B$). In this case, $S(A|B) = 1$ and Alice must send one qubit down the quantum channel to transfer her state to Bob. She could also send half of an EPR state to Bob, and use quantum teleportation⁸ to transfer her state.
- (2) The classically correlated state $\rho_{AB} = \frac{1}{2}(|00\rangle_{AB} + |11\rangle_{AB})$. We imagine this state to be part of a pure state with the reference system R , $|\psi\rangle_{ABR} = \frac{1}{\sqrt{2}}(|0\rangle_A|0\rangle_B|0\rangle_R + |1\rangle_A|1\rangle_B|1\rangle_R)$. In this case, $S(A|B) = 0$, and thus no quantum information needs to be sent. Indeed, Alice can measure her state in the basis $|0\rangle \pm |1\rangle$, and inform Bob of the result. Depending on the outcome of the measurement,

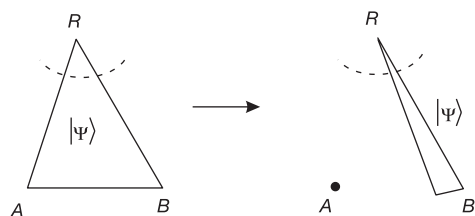


Figure 1 | Diagrammatic representation of the process of state merging. Initially, the state $|\psi\rangle$ is shared between the three systems: R (reference), A (Alice) and B (Bob). After the communication, Alice's system is in a pure state, while Bob holds not only his but also her initial share. Note that the reference's state ρ_R has not changed, as indicated by the curve separating R from AB . The protocol for state merging is as follows: Let Alice and Bob have a large number n of the state ρ_{AB} . To begin, we note that we only need to describe the protocol for negative $S(A|B)$, as otherwise Alice and Bob can share $nS(A|B)$ EPR pairs (by sending this number of quantum bits) and create a state $|\psi\rangle_{AA'BB'R}$ with $S(AA'|BB') < 0$. This is because adding an EPR pair reduces the conditional entropy by one unit. However, $S(A|B) < 0$ is equivalently expressed as $S(A) > S(AB) = S(R)$, and in this case (compare refs 16–18) measurement in a uniformly random basis on Alice's n systems projects Bob and R into a state $|\phi\rangle_{BR}$ whose reduction to R is very close to ρ_R . But this means that Bob can, by a local operation, transform $|\phi\rangle_{BR}$ to $|\psi\rangle_{ABR}$. Finally, by coarse-graining the random measurement, Alice essentially projects onto a good quantum code^{16–18} of rate $-S(A|B)$; this still results in Bob obtaining the full state ρ_{AB} , but now, just under $-nS(A|B)$ EPR pairs are also created. These codes can also be obtained by an alternative construction²³.

Bob and R will share one of two states $|\phi^\pm\rangle_{BR} = \frac{1}{\sqrt{2}}(|0\rangle_B|0\rangle_R \pm |1\rangle_B|1\rangle_R)$, and by a local operation, Bob can always transform the state to $|\psi\rangle_{A'BR} = \frac{1}{\sqrt{2}}(|0\rangle_{A'}|0\rangle_B|0\rangle_R + |1\rangle_{A'}|1\rangle_B|1\rangle_R)$ with A' being an ancilla at Bob's site. Alice has thus managed to send her state to Bob, while fully preserving their entanglement with R .

- (3) For the state $|\phi^+\rangle_{AB} = \frac{1}{\sqrt{2}}(|0\rangle_A|0\rangle_B + |1\rangle_A|1\rangle_B)$, $S(A|B) = -1$, and Alice and Bob can keep this shared EPR pair to allow future transmission of quantum information, whereas Bob creates the EPR pair $|\phi^+\rangle_{A'B}$ locally. That is, transferring a pure state is trivial because the pure state is known and can be created locally.

We now make a couple of observations about state merging. First, the amount of classical communication that is required is given by the number of quantum codes (see description of Fig. 1) in Alice's projection: the quantum mutual information $I(A : R) = S(A) + S(R) - S(AR)$ between Alice and the reference R . Secondly, the measurement of Alice makes her state completely factorized from R , thus reinforcing the interpretation of quantum mutual information as the minimum entropy production of any local decorrelating process^{10,11}. This same quantity is also equal to the amount of irreversibility of a cyclic process: Bob initially has a state, then gives Alice her share (communicating $S(A)$), which is finally merged back to him (communicating $S(A|B)$). The total quantum communication of this cycle is $I(A : R)$ quantum bits.

Because state merging is such a basic primitive, it allows us to solve a number of other problems in quantum information theory fairly easily. We now sketch four particularly striking applications.

Distributed quantum compression: for a single party, a source emitting states with density matrix ρ_A can be compressed at a rate given by the entropy $S(A)$ of the source by performing quantum data compression³. Let us now consider the distributed scenario—we imagine that the source emits states with density matrix $\rho_{A_1A_2\dots A_m}$, and distributes it over m parties. The parties wish to compress their shares as much as possible so that the full state can be reconstructed by a single decoder. Until now, the general solution of this problem has appeared intractable¹², but it becomes very simple once we allow classical side information for free, and use state merging.

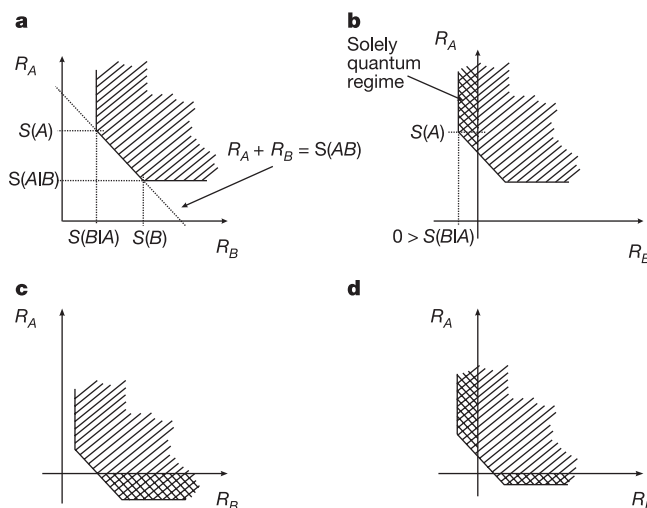


Figure 2 | The rate region for distributed compression by two parties with individual rates R_A and R_B . The total rate R_{AB} is bounded by $S(AB)$. **a**, The rate region of a source with positive conditional entropies; **b**, **c**, the purely quantum case of sources where $S(B|A) < 0$ or $S(A|B) < 0$. It is even possible that both $S(B|A)$ and $S(A|B)$ are negative, as shown in **d**, but observe that the rate-sum $S(AB)$ must be positive. If one party compresses at a rate $S(B)$, then the other party can over-compress at a rate $S(A|B)$, by merging her state with the state that will end up with the decoder. Time-sharing gives the full rate region, since the bounds evidently cannot be improved. Analogously, for m parties A_i , and all subsets $T \subseteq \{1, 2, \dots, m\}$ holding a combined state with the entropy $S(T)$, the rate sums $R_T = \sum_{i \in T} R_{A_i}$ have to obey $R_T \geq S(T|\bar{T})$ with $\bar{T} = \{1, 2, \dots, m\} \setminus T$ the complement of set T .

Remarkably, the parties can compress the state at the total rate $S(A_1 A_2 \dots A_m)$ —the Schumacher limit² for collective compression—even though they must operate separately. This is analogous to the classical result: the Slepian–Wolf theorem³. We describe the quantum solution for two parties and depict the rate region in Fig. 2.

A closely related problem to distributed compression is noiseless coding with side information. Here, only Alice's state needs to arrive at the decoder, while Bob can send part of his state to the decoder in order to help Alice lower her rate. The classical case of this problem was introduced by Wyner¹³. For the quantum case, we require that the full state ρ_{AB} is preserved in the protocol, but do not place any restriction on what part of Bob's state may be at the decoder and what part can remain with him. For one-way protocols, we find by using state merging that if ρ_A and ρ_B are encoded at rates R_A and R_B respectively, then the decoder can recover ρ_A if and only if $R_A \geq S(A|U)$ and $R_B \geq E_p(AU : R) - S(A|U)$. Here R is the purifying reference system, U and V are systems with their joint state produced by acting with some unitary on B and $E_p(AU : R) \equiv \min_A S(AUA(V))$ is the entanglement of purification¹⁴. The minimum is taken over all channels A acting on V .

In addition to the two central questions of information theory we asked earlier—*how much is there to know? how large is our ignorance?*—information theory also concerns itself with communication rates. In the quantum world, the rate at which quantum information can be sent down a noisy channel is related to the coherent information $I(A|B)$ which was previously defined as¹⁵ $\max\{S(B) - S(AB), 0\}$. This quantity is the quantum counterpart of Shannon's mutual information; when $I(A|B)$ is maximized over input states, it gives the rate at which quantum information can be sent from Alice to Bob via a noisy quantum channel^{16–18}. As with the classical conditional entropy, Shannon's classical mutual information¹ is always positive, and indeed it makes no sense to have classical channels with negative capacity. However, the relationship between the coherent information and the quantum channel capacity contained a puzzle. The channel capacity was thought to be meaningful as a positive quantity. So the coherent information was defined as the maximum of 0 and $S(B) - S(AB)$, since it could be negative.

We will see that negative values do make sense, and thus propose that $I(A|B)$ should not be defined as above, but rather as $I(A|B) = -S(A|B)$. It turns out that negative rate, although impossible in classical information theory, has its interpretation in a situation with two senders.

We imagine that Alice and Bob wish to send independent quantum states to a single decoder Charlie via a noisy channel which acts on both inputs. This problem is considered in ref. 19. Our approach using state merging also provides a solution when either of the coherent informations are negative, and gives the following larger achievable rates, compared to ref. 19:

$$R_A \leq I(A|CB), \quad R_B \leq I(B|CA), \quad R_A + R_B \leq I(AB|C) \quad (2)$$

where R_A and R_B are the rates of Alice and Bob for sending quantum states. Here, we use our redefinition of the coherent information, in which we allow it to be negative. In achieving these rates, Alice can send (or invest) $I(A|C)$ quantum bits to merge her state with the decoder. The second sender (Bob) then already has Alice's state at the decoder, and can send at the higher rate $I(B|AC)$. This provides an interpretation of negative achievable rates: if a channel of one sender has negative coherent information, this means that she has to invest this amount of entanglement to help her partner achieve the highest rate. The protocol is for one of the senders to merge his state with the state held by the decoder. The expressions in (2) are in formal analogy with the classical multiple access channel.

Entanglement of assistance: consider Alice, Bob and $m - 2$ other parties sharing (many copies of) a pure quantum state. The entanglement of assistance E_A (ref. 20) is defined as the maximum entanglement that the other parties can create between Alice and Bob by local measurements and classical communication. This was recently

solved for up to four parties²¹, and can be generalized to an arbitrary number m of parties using state merging (using universal codes that depend only on the density matrix of the helper, as described in Fig. 1). We find that the maximal amount of entanglement that can be distilled between Alice and Bob, with the help of the other parties, is given by the minimum entanglement across any bipartite cut of the system which separates Alice from Bob:

$$E_A = \min_T \{S(AT), S(B\bar{T})\} \quad (3)$$

where the minimum is taken over all possible partitions of the other parties into groups T and its complement $\bar{T} = \{1, \dots, m - 2\} \setminus T$. To achieve this, each party in turn merges their state with the remaining parties, preserving the minimum cut entanglement.

We have described a fundamental quantum information primitive—state merging—and demonstrated some of its many applications. There are also conceptual implications. For example, the celebrated strong subadditivity of quantum entropy²², $S(A|BC) \leq S(A|B)$, receives a clear interpretation and transparent proof: having more prior information makes state merging cheaper. Our results also shed new light on the foundations of quantum mechanics: it has long been known that there are no conditional probabilities, so defining conditional entropy is problematic. Merely replacing classical entropy with quantum entropy gives a quantity which can be positive or negative. Quite paradoxically, only the negative part was understood operationally—as quantum channel capacity—which made the problem even more obscure. State merging ‘annihilates’ these problems with each other. It turns out that the puzzling form of quantum capacity as a conditional entropy is the flip side of our interpretation of quantum conditional entropy as partial quantum information, which makes equal sense in the positive and negative regime. The key point is to realize that in the negative regime, we can gain entanglement and transfer Alice's partial state, while in the positive regime, only the partial state is transferred.

Remarkably, and despite the formal analogy, the classical scenario does not occur as a classical limit of the quantum scenario—we consider both classical and quantum communication, and there is no meaning to preserving entanglement in the classical case. We have only just begun to grasp the full implications of state merging and negative partial information; a longer technical account (M.H., J.O. and A.W., manuscript in preparation), with rigorous proofs and further application will appear elsewhere.

Received 20 February; accepted 9 June 2005.

- Shannon, C. E. A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 379–423, 623–656 (1948).
- Schumacher, B. Quantum coding. *Phys. Rev. A* **51**, 2738–2747 (1995).
- Slepian, D. & Wolf, J. K. Noiseless coding of correlated information sources. *IEEE Trans. Inf. Theory* **19**, 461–480 (1971).
- Wehrl, A. General properties of entropy. *Rev. Mod. Phys.* **50**, 221–260 (1978).
- Horodecki, R. & Horodecki, P. Quantum redundancies and local realism. *Phys. Lett. A* **194**, 147–152 (1994).
- Cerf, N. J. & Adami, C. Negative entropy and information in quantum mechanics. *Phys. Rev. Lett.* **79**, 5194–5197 (1997).
- Einstein, A., Podolsky, B. & Rosen, N. Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev.* **47**, 777–780 (1935).
- Bennett, C. H. *et al.* Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Phys. Rev. Lett.* **70**, 1895–1899 (1993).
- Schrödinger, E. Die gegenwärtige Situation der Quantenmechanik. *Naturwissenschaften* **23**, 807–821, 823–828, 844–849 (1935).
- Groisman, B., Popescu, S. & Winter, A. On the quantum, classical and total amount of correlations in a quantum state. Preprint at (<http://arXiv.org/quant-ph/0410091>) (2004).
- Horodecki, M. *et al.* Local versus non-local information in quantum information theory: formalism and phenomena. *Phys. Rev. A* **71**, 062307 (2005).
- Ahn, C., Doherty, A., Hayden, P. & Winter, A. On the distributed compression of quantum information. Preprint at (<http://arXiv.org/quant-ph/0403042>) (2004).
- Wyner, A. On source coding with the side information at the decoder. *IEEE Trans. Inf. Theory* **21**, 294–300 (1975).
- Terhal, B. M., Horodecki, M., DiVincenzo, D. P. & Leung, D. The entanglement

- of purification. *J. Math. Phys.* **43**, 4286–4298 (2002).
15. Schumacher, B. & Nielsen, M. A. Quantum data processing and error correction. *Phys. Rev. A* **54**, 2629–2635 (1996).
 16. Shor, P. The quantum channel capacity and coherent information. *MSRI Workshop on Quantum Computation* (Berkeley, 2002); lecture notes available at (<http://www.msri.org/publications/In/msri/2002/quantumcrypto/shor/1>).
 17. Lloyd, S. The capacity of the noisy quantum channel. *Phys. Rev. A* **55**, 1613–1622 (1997).
 18. Devetak, I. The private classical capacity and quantum capacity of a quantum channel. *IEEE Trans. Inf. Theory* **51**(1), 44–55 (2005).
 19. Yard, J., Devetak, I. & Hayden, P. Capacity theorems for quantum multiple access channels—Part I: classical-quantum and quantum-quantum capacity regions (2005). Preprint at (<http://arXiv.org/quant-ph/0501045>) (2005).
 20. DiVincenzo, D. P., et al. in *Proc. 1st NASA Int. Conf. on Quantum Computing and Quantum Communication* (ed. Williams, C. P.) LNCS 1509 247–257 (Springer, London, 1999).
 21. Smolin, J. A., Verstraete, F. & Winter, A. Entanglement of assistance and multipartite state distillation. Preprint at (<http://arXiv.org/quant-ph/0505038>) (2005).
 22. Lieb, E. H. & Ruskai, M. B. Proof of the strong subadditivity of quantum-mechanical entropy with an appendix by B. Simon. *J. Math. Phys.* **14**, 1938–1941 (1973).
 23. Devetak, I. & Winter, A. Relating quantum privacy and quantum coherence: an operational approach. *Phys. Rev. Lett.* **93**, 080501 (2004).
- Acknowledgements** This work was done at the Isaac Newton Institute (Cambridge) in August–December 2004 and we are grateful for the institute's hospitality. We thank W. Unruh for comments on an earlier draft of this paper. We acknowledge the support of EC grants RESQ, QUPRODIS and PROSECCO. M.H. was additionally supported by the Polish Ministry of Scientific Research and Information Technology, J.O. by the Cambridge-MIT Institute and Newton Trust, and A.W. by EPSRC's "IRC QIP".
- Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.O. (j.oppenheim@damtp.cam.ac.uk).

Measurement of the conductance of single conjugated molecules

Tali Dadoosh^{1,2*}, Yoav Gordin^{1*}, Roman Krahne^{1†}, Ilya Khivrich¹, Diana Mahalu¹, Veronica Frydman³, Joseph Spering², Amir Yacoby¹ & Israel Bar-Joseph¹

Electrical conduction through molecules depends critically on the delocalization of the molecular electronic orbitals and their connection to the metallic contacts. Thiolated (–SH) conjugated organic molecules are therefore considered good candidates for molecular conductors^{1,2}: in such molecules, the orbitals are delocalized throughout the molecular backbone, with substantial weight on the sulphur–metal bonds^{1–4}. However, their relatively small size, typically ~1 nm, calls for innovative approaches to realize a functioning single-molecule device^{5–11}. Here we report an approach for contacting a single molecule, and use it to study the effect of localizing groups within a conjugated molecule on the electrical conduction. Our method is based on synthesizing a dimer structure, consisting of two colloidal gold particles connected by a dithiolated short organic molecule^{12,13}, and electrostatically trapping it between two metal electrodes. We study the electrical conduction through three short organic molecules: 4,4'-biphenyldithiol (BPD), a fully conjugated molecule; bis-(4-mercaptophenyl)-ether (BPE)¹⁴, in which the conjugation is broken at the centre by an oxygen atom; and 1,4-benzenedimethanethiol (BDMT), in which the conjugation is broken near the contacts by a methylene group. We find that the oxygen in BPE and the methylene groups in BDMT both suppress the electrical conduction relative to that in BPD.

Various methods have been suggested and realized to form nanometre-sized gaps, including electron beam lithography with shadow mask evaporation¹⁵, mechanical break junctions^{5–7}, electromigration⁸, and side etched quantum wells¹⁶. Molecules are then deposited on the thin gap, and it is assumed that electrical conductivity is obtained when a single molecule bridges the gap and conducts current. Although impressive success in measuring the conductance of molecular junctions has been reported^{5–7,9,10,17}, serious problems are evident. The most notable are the uncertainty about the number of conducting molecules in the junction, and the lack of information about the shape and structure of the metal contacts near the molecule. The dimer-based contacting scheme presented here provides several advantages. (1) Single-molecule devices can be fabricated with high certainty. (2) The contacts to the molecule are well defined and can be characterized separately. (3) Our scheme avoids the need for fabricating nanometre-sized gaps. (4) It also allows measurements of the temperature dependence of the molecular junction conductance over periods of hours and even days, regaining the original spectra after thermal cycling.

The dimer synthesis is performed by mixing a solution of dithiolated molecules, whose structure is depicted in Fig. 1a, with a gold colloid, keeping their respective ratio below 1:10. The colloids were stabilized using the conventional citrate method¹⁸. In making the bridged dimers, the thiols on the bridging molecule displace citrate

anions to form stable Au–S bonds¹⁹ (Fig. 1b, c). If more than one molecule binds to a certain colloidal particle, trimers (Fig. 1d), tetramers (Fig. 1e) and higher oligomers can be formed. To ensure that most of the dimeric colloidal particles are bridged by a single dithiolated molecule, the concentration of the latter (C_m) in the reaction mixture should be much smaller than the concentration of the monomeric colloidal particles (C_p). In that case, when the reaction is completed, the ratio R of dimers to single unbound colloidal particles is expected to follow the relation $R \approx C_m/(C_p - 2C_m)$. We measured R for different input concentrations of dithiolated molecules, keeping C_p fixed, and found that the above relation holds for $C_m:C_p$ ratios that are lower than 1:10.

The dimeric colloidal particles are separated from monomeric particles and higher oligomers on the basis of their relative mass, by centrifugation through glycerol or sucrose density gradients. Each structure propagates as a different band along the centrifuged tube, and is easily distinguished because of its reddish colour. Figure 1f shows a transmission electron microscope (TEM) image at higher magnification of a dimer (made of two 10-nm colloidal gold particles connected by BDMT) that was synthesized in this manner. The gap between the two particles is readily seen, and its size is in good agreement with the size of the BDMT molecule, which is

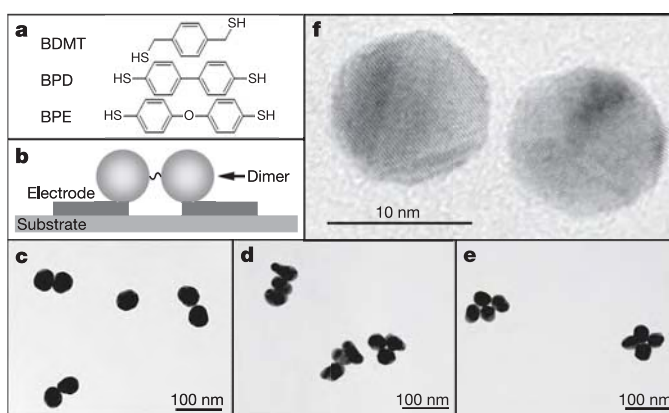


Figure 1 | The molecules and colloidal structures under study. a, The structures of the three molecules: 1,4-benzenedimethanethiol (BDMT), 4,4'-biphenyldithiol (BPD) and bis-(4-mercaptophenyl)-ether (BPE). **b**, The dimer contacting scheme. **c–e**, TEM images of BDMT dimer, trimer and tetramer structures made of 50-nm colloidal gold particles. **f**, TEM image of a BDMT dimer made of 10-nm colloidal gold particles. The ~1 nm separation between the two particles corresponds approximately to the BDMT length (0.9 nm).

¹Department of Condensed Matter Physics, ²Department of Organic Chemistry, ³Chemical Research Support, Weizmann Institute of Science, 76100 Rehovot, Israel.

[†]Present address: National Nanotechnology Labs of INFN, Lecce 73100, Italy.

*These authors contributed equally to this work.

approximately 0.9 nm. To further establish the presence of a molecule in the dimer, we conducted Raman scattering measurements on dimers made from colloidal silver particles²⁰. A clear Raman spectrum was detected from a single dimer, with the characteristic temporal blinking of a single molecule.

We position the dimer between two electrodes using the method of electrostatic trapping^{15,16,21}. The gold electrodes are defined on a silicon substrate with a thin (25–100 nm) oxide layer using electron beam lithography. In most of our studies we use dimers of 30-nm-diameter colloidal gold particles. Hence, the electrodes are made with 40–50 nm separation, which is below the dimer size and larger than the diameter of a single particle. The processed structures are first cleaned²² and then inserted in a probe station where the electrostatic trapping is performed.

The success rate of the electrostatic trapping is relatively high, and reaches ~50% when the impedance of the dimer solution is high enough, such that the dipole–dipole interaction between the electrodes and the colloids is not screened. An analysis of the trapping events reveals that in 55% of the cases the gap is bridged by one colloidal particle only. In 45% of the cases we get a dimer between the two electrodes (Fig. 2a): in 35% the conductivity is too low to determine the conductance features. We find in scanning electron microscopy (SEM) imaging that the brightness of the colloidal gold particles of these dimers is low, probably due to poor conductivity between one of the particles and the electrode²³. In 10% of the trapping events we obtain high conductivity dimers. These are the samples that we refer to here: 11 BPD, 13 BPE and 15 BDMT devices.

Trapping single colloidal particles between closely spaced electrodes (25–30 nm) allows us to study the current–voltage characteristics of the contacts. For the 30-nm colloidal gold particles, we find a Coulomb blockade staircase with a single electron charging energy of ~20–30 meV, and by applying a voltage to the substrate we can measure the diamond structure of the conductance (see Supplementary Fig. 1) characteristic of a single-electron transistor²⁴. It is evident that a tunnel barrier with a typical resistance of ~10 M Ω is formed between the gold particles and the electrodes, probably due to the ligands that stabilize the colloids in the solution.

The measurements of electrical transport through the dimer structures are performed using an a.c. voltage with an amplitude of a few millivolts superimposed on a d.c. voltage, and the current as well as the differential conductance (dI/dV) are measured

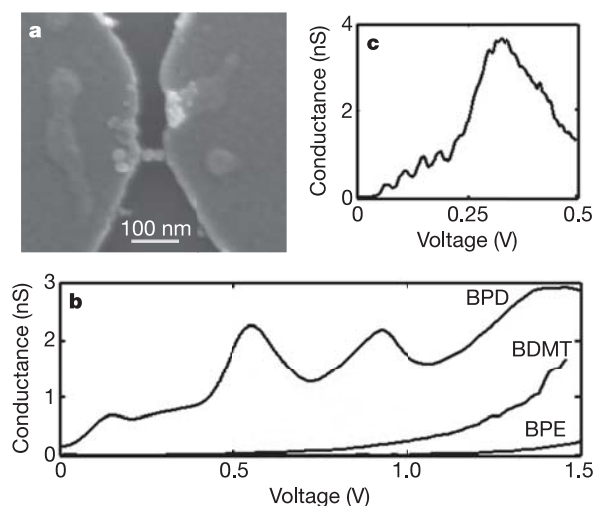


Figure 2 | Image and low-temperature differential conductance spectra of dimers. **a**, SEM image of a dimer trapped between the two electrodes. **b**, The differential conductance as a function of voltage measured for BPD, BDMT and BPE dimers at 4.2 K. **c**, Coulomb blockade oscillations of the colloidal gold particle which are superimposed on the conduction peak of BPD measured at 4.2 K.

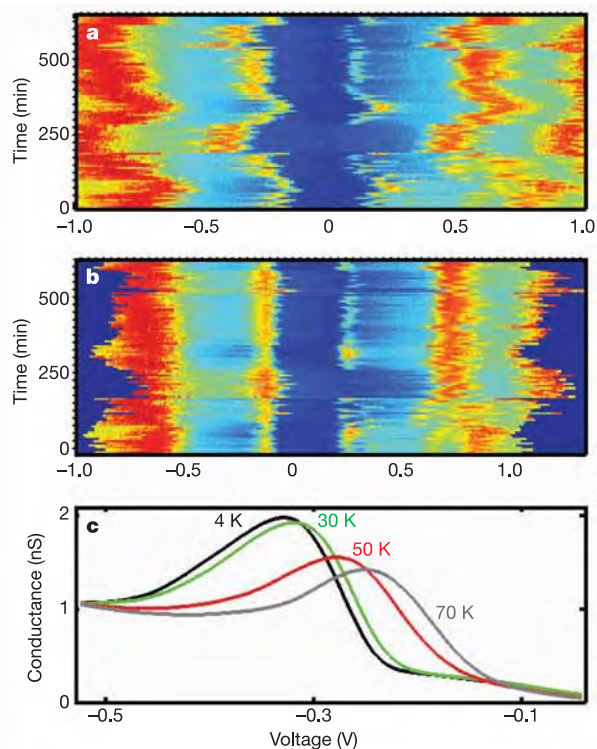


Figure 3 | Temporal fluctuations and temperature dependence of the dimer conductance. **a**, A compilation of 130 measurements of the differential conductance spectrum of one BPD dimer, taken over a period of a few hours. The conductance values are represented by a colour scale, which ranges from blue (0 nS) to red (3 nS). **b**, The spectra of **a** shifted such that their peaks at -0.2 V coincide, thus demonstrating the rigid shift nature of the temporal fluctuations. **c**, The evolution of the lowest voltage conductance peak of a BPD dimer with temperature. A decrease in height and broadening of the peak with increasing temperature are observed.

simultaneously. Figure 2b shows a comparison of the differential conductance of the three molecules at 4 K. These are representative curves from more than ten devices of each of the molecular species that were studied. The BPD dimers clearly show a series of conductance peaks. These peaks cannot be attributed to charging of the gold particles—their typical spacing is a few hundred millivolts, more than an order of magnitude larger than the charging energy of the particles. In fact, in some cases we could resolve the Coulomb blockade oscillations superimposed on the large conduction peaks (see Fig. 2c). The oscillation period is very similar to that observed in a single-particle measurement (see comparison in Supplementary Information). We note that the conductance through a single trapped colloidal particle is substantially higher than that of a dimer, typically by one to two orders of magnitude. Hence, the tunnel barriers between the metal electrodes and the gold particles act as small series resistances, carrying only a minute fraction of the applied voltage, while most of the voltage is dropped across the molecule.

The BPE and BDMT dimers show remarkably different behaviours. It is clearly seen that the conduction through the dimers formed from these molecules turns on at much higher voltages (typically larger than 0.5 V) and one cannot resolve any peak structure. The systematic difference between the three molecules further confirms that the measured differential conductance is of molecular origin. The apparent gap and exponential turn-on of the differential conductance in the case of the BPE and BDMT dimers suggest that adding localizing groups interferes with the conjugated aromatic system and suppresses the overall conductance through the molecule. This assertion is supported by comparing the BPE with the BPD molecule, where the addition of an oxygen atom between

the conjugated rings suppresses the conductance almost entirely below 1 V. A similar effect takes place in the BDMT dimers, where the methylene groups suppress the overlap of the molecular backbone orbitals with the contacts²³.

Focusing on the peak structure of the BPD dimers, we find that while the main features of the conductance spectrum are highly reproducible, there are temporal fluctuations as well as variations between different devices. Figure 3a shows a compilation of 130 measurements of the conductance of one dimer taken over a period of a few hours. Strong temporal fluctuations in peak position, which could be as large as a few hundred millivolts, are evident. We find that the conductance spectrum shifts rigidly—that is, all peaks at positive and negative voltages move in the same direction by approximately the same value. This is demonstrated in Fig. 3b, where we shift all spectra such that their peaks at -0.2 V coincide. It is seen that this operation causes all other peaks to be aligned as well. Such behaviour is typical of electrostatically gated Coulomb blockaded systems with different input and exit capacitances²⁴. Plotting the conductance in the V_{ds} – V_g plane (where V_{ds} and V_g are the drain–source and gate voltages, respectively) produces a skewed diamonds structure (see Supplementary Fig. 1b). It can be seen that under these conditions the conductance peaks at positive and negative drain–source bias indeed move rigidly on application of a gate voltage. As the gold particles are very effective in screening the effect of remote potentials, we conclude that charge movement in close vicinity to the molecule gives rise to the observed rigid shift of the spectrum. These temporal fluctuations can therefore be considered as evidence for gating of the molecule, which occurs randomly. We note that the temporal frequency of this random gating depends on the voltage sweep rate, dV/dt , and the voltage range of the measurement. It could be substantially minimized by working at low sweep rates and within a limited voltage range, such that stable and reproducible measurements can be conducted over hours.

The dimer structure allows us to conveniently perform measurements of the temperature dependence of the molecular junction conductance. We demonstrate this capability in Fig. 3c, which shows the behaviour of the lowest voltage peak (~ -0.3 V in Fig. 3a) as the temperature is increased from 4 K to 100 K. It is evident that the peak broadens and decreases in height with increasing temperature, keeping the area approximately constant. The shift in peak position between the low- and high-temperature measurements is due

to gating, which occurred during this long measurement. The broadening of the peak can be accounted for by the change in thermal distribution of the electrons in the leads. This mechanism is expected to broaden the peak by $\sim 2k_B T$ (half-width at half-maximum), where k_B is the Boltzmann constant.

Comparing the conductance of nine BPD dimer-based devices shows that the variations between devices in the positions and relative strengths of the peaks cannot be explained only as a rigid shift. This is demonstrated in Fig. 4a, which compares the spectra of two BPD dimers. We can see that while the main features appear at approximately the same voltages, one cannot align the two spectra by a simple shift operation. We note also that there are differences in the value of the conductance between the various devices, which can be as large as an order of magnitude. To statistically quantify this variability, we construct a histogram of peak positions (Fig. 4b). This is done by converting each spectrum to a set of peaks, each represented by a block of unit height, a width which is the full-width at half-maximum, and is centred around its maximum. A rigid shift of each spectrum was allowed to accommodate the temporal fluctuations. It is seen that the peak structure survives the averaging over many molecules, indicating that the variability in peak position is smaller than the peak spacing. The peak positions determined from the histogram agree very well with measurements done on self-assembled monolayers of BPD²⁶. However, such agreement is lacking when comparing the measured spectra with recent calculations of molecular conduction^{3,4}. In particular, the measured turn-on voltage is significantly lower than theoretically predicted. Moreover, the predicted conductance is several orders of magnitude higher than observed.

METHODS

Colloidal gold nanoparticles with a diameter of 30 nm were prepared using a method described by Grabar *et al.*¹⁸, with slight modifications. Briefly, a solution of 'seed colloid' (2.6 nm) was prepared by adding 1.0 ml of 10 mg ml^{-1} aqueous $\text{HAuCl}_4 \cdot 3\text{H}_2\text{O}$ to 100 ml of Milli-Q water while stirring. One min later, 1.0 ml of 10 mg ml^{-1} aqueous trisodium citrate was added, followed, 1 min later, by 1.0 ml of freshly prepared 0.75 mg ml^{-1} NaBH_4 in water, and the mixture was stirred for 5 min. Larger colloidal particles, 30 nm in diameter, were prepared by adding 0.2 ml of 10 mg ml^{-1} trisodium citrate and $130 \mu\text{l}$ of the Au 'seed colloid' solution to 50 ml of a boiling aqueous solution of $\text{HAuCl}_4 \cdot 3\text{H}_2\text{O}$ (0.1 mg ml^{-1}) while stirring. Refluxing was continued for 10 min, or until a change to a reddish colour was observed. This procedure resulted in a citrate-stabilized colloid of gold particles well-dispersed in water²⁷; the particles had a narrow size distribution²⁸, as confirmed by TEM. The concentration of colloidal gold particles was calculated on the basis of mean particle diameter measured by TEM and the weight of HAuCl_4 used in the reaction¹⁹.

The dimer preparation consisted of several stages. First, the colloidal gold solution was concentrated. This was achieved by centrifuging the solution at $9.7 \times 10^3 g$ for 5 min and discarding the supernatant. The precipitate was resuspended in 2.5 ml of 1 mM NaOH and centrifuged again as above, leaving about 0.4 ml of concentrated solution. The spacer dithiol molecule was dissolved in 0.1 M NaOH and diluted with 1 mM NaOH to the desired concentration. This solution was mixed with an equal volume of the concentrated colloidal gold solution containing > 10 -fold molar excess of colloidal particles and incubated for 24 h at 4 °C. The mixture (0.5 ml) was loaded on a 10–75% glycerol gradient in water and centrifuged at $1 \times 10^4 g$ for 10 min in a Beckman SW41 rotor. Fractions (0.5 ml) of the gradient were collected and characterized by TEM.

The electrodes are prepared on an electron beam defined pattern, and consist of 20 nm of Au on 10 nm of Ni that serves as an adhesion layer. The electrodes are cleaned in warm acetone and methanol for 5 min each, then placed in an ozone stripper for 8 min and finally stirred in ethanol for 10 min (ref. 22).

Electrostatic trapping is performed in a probe station; a $0.35 \mu\text{l}$ drop of the dimer solution (containing typically 45–55% dimers, 40–50% single particles and a few per cent of higher conjugates) is placed on the electrodes' surface. Next an a.c. voltage of 800 mV at 10 MHz is applied between the electrodes for 1 min. The sample is then cleaned in double distilled water, and blown dry in nitrogen. The devices are tested on the probe station, and the ones that show conductivity are bonded and measured at 4 K. After the measurement, the devices are imaged in an SEM to determine the type of colloidal structure that has been trapped.

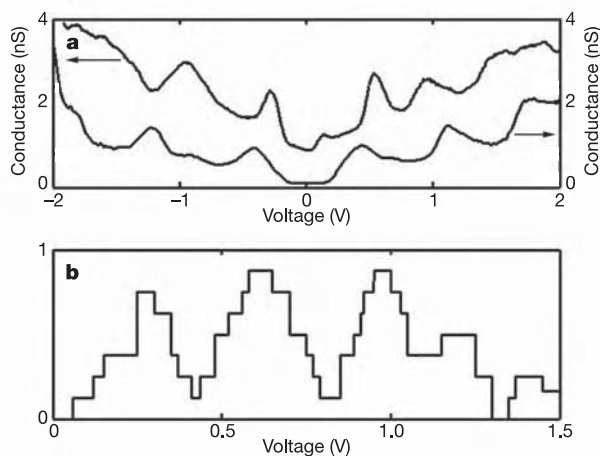


Figure 4 | The reproducibility of the conductance spectra measurements of different BPD dimers. **a**, Comparison of the differential conductance spectrum of two different BPD devices (the upper curve has been shifted upwards for clarity). **b**, Histogram of the conductance peak positions collected from nine BPD devices. Each spectrum was converted to a set of peaks of unit height and a width which is the full-width at half-maximum. A rigid shift of the spectrum by up to 50 mV was allowed.

Received 20 January; accepted 7 June 2005.

- Nitzan, A. & Ratner, M. A. Electron transport in molecular wire junctions. *Science* **300**, 1384–1389 (2003).
- Joachim, C., Gimzewski, J. K. & Aviram, A. Electronics using hybrid-molecular and mono-molecular devices. *Nature* **408**, 541–548 (2000).
- Xue, Y. & Ratner, M. A. Microscopic study of electrical transport through individual molecules with metallic contacts. I. Band lineup, voltage drop, and high-field transport. *Phys. Rev. B* **68**, 115406–115418 (2003).
- Remacle, F. & Levine, R. D. Electrical transmission of molecular bridges. *Chem. Phys. Lett.* **383**, 537–543 (2004).
- Reed, M. A., Zhou, C., Muller, C. J., Burgin, T. P. & Tour, J. M. Conductance of a molecular junction. *Science* **278**, 252–254 (1997).
- Reichert, J. *et al.* Driving current through single organic molecules. *Phys. Rev. Lett.* **88**, 176804 (2002).
- Smit, R. H. M. *et al.* Measurement of the conductance of a hydrogen molecule. *Nature* **419**, 906–909 (2002).
- Park, H., Lim, A. K. L., Alivisatos, A. P., Park, J. & McEuen, P. L. Fabrication of metallic electrodes with nanometer separation by electromigration. *Appl. Phys. Lett.* **75**, 301–303 (1999).
- Park, J. *et al.* Coulomb blockade and the Kondo effect in single-atom transistors. *Nature* **417**, 722–725 (2002).
- Liang, W., Shores, M. P., Bockrath, M., Long, J. R. & Park, H. Kondo resonance in a single-molecule transistor. *Nature* **417**, 725–729 (2002).
- Cui, X. D. *et al.* Reproducible measurement of single-molecule conductivity. *Science* **294**, 571–574 (2001).
- Peng, X., Wilson, T. E., Alivisatos, A. P. & Schultz, P. G. Synthesis and isolation of a homodimer of cadmium selenide nanocrystals. *Angew. Chem. Int. Edn Engl.* **36**, 145–147 (1997).
- Brousseau, L. C. III, Novak, J. P., Marinakos, S. M. & Feldheim, D. L. Assembly of phenylacetylene-bridged gold nanocluster dimers and trimers. *Adv. Mater.* **11**, 447–449 (1999).
- Baron, A. L. & Blank, D. R. Synthesis and properties of some aromatic polythioethers. *Makromol. Chem.* **140**, 83–89 (1970).
- Bezryadin, A., Dekker, C. & Schmid, G. Electrostatic trapping of single conducting nanoparticles between nanoelectrodes. *Appl. Phys. Lett.* **71**, 1273–1275 (1997).
- Krahne, R. *et al.* Fabrication of nanoscale gaps in integrated circuits. *Appl. Phys. Lett.* **81**, 730–732 (2002).
- Park, H. *et al.* Nanomechanical oscillations in a single-C₆₀ transistor. *Nature* **407**, 57–60 (2000).
- Grabar, K. C. *et al.* Two-dimensional arrays of colloidal gold particles: A flexible approach to macroscopic metal surfaces. *Langmuir* **12**, 2353–2361 (1996).
- Weisbecker, C. S., Merritt, M. V. & Whitesides, G. M. Molecular self-assembly of aliphatic thiols on gold colloids. *Langmuir* **12**, 3763–3772 (1996).
- Xu, H. X., Bjerneld, E. J., Kall, M. & Borjesson, L. Spectroscopy of single hemoglobin molecules by surface enhanced Raman scattering. *Phys. Rev. Lett.* **83**, 4357–4360 (1999).
- Amlani, I., Rawlett, A. M., Nagahara, L. A. & Tsui, R. K. An approach to transport measurements of electronic molecules. *Appl. Phys. Lett.* **80**, 2761–2763 (2002).
- Ron, H., Matlis, S. & Rubinstein, I. Self-assembled monolayers on oxidized metals. 2. Gold surface oxidative pretreatment, monolayer properties, and depression formation. *Langmuir* **14**, 1116–1121 (1998).
- Hazani, M. *et al.* DNA-mediated self-assembly of carbon nanotube-based electronic devices. *Chem. Phys. Lett.* **391**, 389–392 (2004).
- Grabert, H. & Devoret, M. H. (eds) *Single Charge Tunneling: Coulomb Blockade Phenomena in Nanostructures* (NATO ASI Ser. B. Vol. 294, Plenum, New York, 1992).
- Xiao, X. Y., Xu, B. Q. & Tao, N. J. Measurement of single molecule conductance: Benzenedithiol and benzenedimethanethiol. *Nano Lett.* **4**, 267–271 (2004).
- Lee, J. O. *et al.* Absence of strong gate effects in electrical measurements on phenylene-based conjugated molecules. *Nano Lett.* **3**, 113–117 (2003).
- Zhu, T., Vasilev, K., Kreiter, M., Mittler, S. & Knoll, W. Surface modification of citrate-reduced colloidal gold nanoparticles with 2-mercaptosuccinic acid. *Langmuir* **19**, 9518–9525 (2003).
- Handley, D. A. in *Colloidal Gold—Principles, Methods, and Applications* (ed. Hayat, M. A.) 13–32 (Academic, New York, 1989).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the Minerva foundation. We thank R. Popovitz for assistance in obtaining the TEM image (Fig. 1f).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to I.B.-J. (hbar@wisemail.weizmann.ac.il).

Stability of the Larsen B ice shelf on the Antarctic Peninsula during the Holocene epoch

Eugene Domack¹, Diana Duran¹, Amy Leventer², Scott Ishman³, Sarah Doane¹, Scott McCallum³, David Amblas⁴, Jim Ring⁵, Robert Gilbert⁶ & Michael Prentice⁷

The stability of the Antarctic ice shelves in a warming climate has long been discussed¹, and the recent collapse of a significant part, over 12,500 km² in area, of the Larsen ice shelf off the Antarctic Peninsula^{2,3} has led to a refocus toward the implications of ice shelf decay for the stability of Antarctica's grounded ice^{4–6}. Some smaller Antarctic ice shelves have undergone periodic growth and decay over the past 11,000 yr (refs 7–11), but these ice shelves are at the climatic limit of ice shelf viability¹² and are therefore expected to respond rapidly to natural climate variability at century to millennial scales^{8–11}. Here we use records of diatoms, detrital material and geochemical parameters from six marine sediment cores in the vicinity of the Larsen ice shelf to demonstrate that the recent collapse of the Larsen B ice shelf is unprecedented during the Holocene. We infer from our oxygen isotope measurements in planktonic foraminifera that the Larsen B ice shelf has been thinning throughout the Holocene, and we suggest that the recent prolonged period of warming in the Antarctic Peninsula region^{13,14}, in combination with the long-term thinning, has led to collapse of the ice shelf.

A marine geologic survey was conducted in the northwestern Weddell Sea (between 65°S and 66°S, and 59°W and 61°W) just before the dramatic collapse of a large portion of the Larsen ice shelf B (LIS-B; colloquially Larsen B ice shelf) in March 2002 (Fig. 1). Here we collected sediment cores, surface sediment grabs, bottom photographs and oceanographic profiles of temperature and salinity, and conducted a reconnaissance bathymetric survey (Figs 2 and 3). These data were collected in open water just east of the pre-collapse LIS-B front, but following a series of calving events resulting in shelf edge retreat before the 2002 collapse. Thus the LIS-B cores and an additional core¹¹ previously collected in the LIS-A region (core 23 in Fig. 1a) were located in a sub-ice shelf setting until 1995. Surface phytoplankton concentrations were estimated via ocean colour satellite imagery (Fig. 1c) and surface water sampling. The sediment cores recovered glacial and Holocene marine sedimentary sequences representing the transitions from glacial conditions to de-glacial (sub-ice shelf) conditions, and then to open marine conditions. Three stratigraphic intervals (units 3, 2 and 1) are recognized: a lower homogeneous diamicton (unit 3; glacial), a stratified to cross-stratified gravelly sand to granulated mud (unit 2; deglacial), and an upper laminated to homogeneous, slightly sandy mud (unit 1; sub-ice shelf) (Fig. 2b). Only the uppermost centimetre of unit 1, with higher biosiliceous content, represents open marine conditions. These units are interpreted on the basis of a number of specific criteria in Antarctic glacial marine sequences^{11,15,16}. The presence of glacial till (unit 3) is consistent with the glacially sculpted and streamlined sea floor documented from multibeam data (Fig. 1b)

and with a much expanded glacial regime on both sides of the Antarctic Peninsula (Fig. 1a; see Supplementary Information).

Termination of glacial cover upon the sea floor of both LIS-A and LIS-B is constrained by radiocarbon dates of $10,600 \pm 55$ years before present (yr BP) on calcareous foraminifera within the base of the sub-ice shelf facies, unit 1 in core KC-2 (Table 1), and palaeomagnetic intensity cross-dates of $10,700 \pm 500$ yr BP within LIS-A cores¹¹. Radiocarbon ages on calcareous foraminifera and organic matter (Table 1) demonstrate that subsequent deposition beneath the LIS-B took place at progressively lower rates following deglaciation. In support of lithofacies interpretations is the twofold greater thickness for unit 1 within core KC-3, when compared to other cores (Fig. 2b). KC-3 is closest to the historical grounding line of the LIS-B around Jason Peninsula (Fig. 1b) and would be expected to have received glacial detritus at greater rates than sites more distal from the debris source (the grounding line).

All of the cores except KC-3 are characterized by concentrations of gravel at the sediment–water interface (Fig. 2a) marking the accumulation of coarse ice-rafted material. This is a result of a combination of coarse debris released when the ice shelf retreated to its penultimate position in 1998¹⁵ and the rafting of poorly sorted stones under the distal reaches of the LIS-B, which was generally devoid of material except for englacial debris septa entrained at tributary junctures along the Antarctic Peninsula¹⁷. The inference that some of these stones resided exposed on the sea floor for some time is supported by the MnO₂ coatings of a few micrometres thickness over their surface, requiring at least 1.5–2.7 kyr to accumulate given observed rates of MnO₂ accretion in the Southern Ocean¹⁸. The surface of unit 1 is not a winnowed or lag surface but one of low sedimentation (Table 1) as ²¹⁰Pb activities show near-surface excess beyond background levels (Fig. 3b), albeit associated with very low sediment accumulation rates (Table 1). Only a few of the large boulders recovered and observed in bottom photographs (Fig. 2b) show evidence of recent colonization by epifaunal organisms. All such organisms yield modern ¹⁴C ages for biogenic calcite (Table 1, AA-49353 to AA-49358), attesting to the recent increase in advection of organic detritus to the sea floor in this area. Extremely low diatom abundance within all except the upper centimetre of unit 1 supports a sub-ice shelf origin for unit 1.

Quantitative downcore diatom data (Fig. 3a) indicate contrasting Holocene histories for LIS-A and LIS-B. It is notable that there is a much higher overall diatom concentration in KC23 (LIS-A) as contrasted to concentrations in LIS-B cores, which are several orders of magnitude lower, reinforcing the hypothesis that while the LIS-A region had no ice shelf for extended periods during the Holocene¹¹, the LIS-B persisted through the Holocene until its recent collapse. A significant increase in absolute diatom concentration noted in some

¹Department of Geosciences, Hamilton College, Clinton, New York 13323, USA. ²Department of Geology, Colgate University, Hamilton, New York 13346, USA. ³Department of Geology, Southern Illinois University, Carbondale, Illinois 62901, USA. ⁴Department of Stratigraphy, Paleontology, and Marine Geosciences, University of Barcelona, 08028 Barcelona, Spain. ⁵Physics Department, Hamilton College, Clinton, New York 13323, USA. ⁶Department of Geography, Queen's University, Kingston, Ontario ON K7L 3N6, Canada. ⁷Department of Earth Science, University of New Hampshire, Durham, New Hampshire 03824, USA.

surface sediments from LIS-B illustrates the impact of very recent ice shelf breakout on underlying sediments (Fig. 3a). The heterogeneous nature of the sea-floor surface (Fig. 2a) and the short duration over which open water has existed in this area account for the heterogeneity of surface diatom peaks. These modern values are orders of magnitude lower than values recorded from other parts of the western Weddell Sea shelf¹⁹, reflecting relatively low annual production in a region still characterized by high seasonal sea ice concentration. Surface sediment diatom assemblages, including *Fragilariopsis curta*, document an environment dominated by sea ice. The valves are lightly silicified; reinforcing that dissolution is not the principal reason for low diatom abundances in LIS-B surface sediments.

Apart from the upper centimetre, sparse diatom valves and fragments occur within unit 1, attributed to advection and deposition under the ice shelf, as observed under other ice shelves²⁰. Phytoplankton and surface chlorophyll concentrations from satellite data demonstrate a large algal bloom east of the ice front which, together with the observed bottom current trajectories (Figs 1c and 2a), support the potential for advection of biogenic components underneath the ice shelf. The distinct increase in foraminiferal abundance in unit 1 is paralleled by diatom data;

while low concentrations of diatoms are found in unit 1, even diatom fragments are nearly absent in unit 2, and are indeed completely lacking in unit 3.

Finally, the petrologic composition of terrigenous grains within the sediment cores indicate only subtle shifts in debris provenance consistent with sub-glacial to sub-ice shelf deposition, and a source limited to substrates eroded upstream of glacial or ice shelf flow^{10,16,21} (Supplementary Information). Exotic iceberg rafted detritus is a key characteristic in recognition of open marine versus sub-ice shelf palaeoenvironments^{10,11,16}.

All the evidence indicates that LIS-B has been a stable component of the northwestern Weddell Sea since the Late Pleistocene to Holocene transition, 11.5 kyr BP. In contrast, the northernmost portions of a much larger Larsen ice shelf system and small ice shelves on the western side of the Antarctic Peninsula experienced periods of decay and open marine conditions during the Holocene^{7–11,15}. Our observation, that the modern collapse of the LIS-B is a unique event within the Holocene, supports the hypothesis that the current warming trend in the northwestern Weddell Sea has exceeded past warm episodes in both its magnitude and duration.

The suggestion that the LIS has recently experienced accelerated thinning due to under-melt²² may also help explain the unique

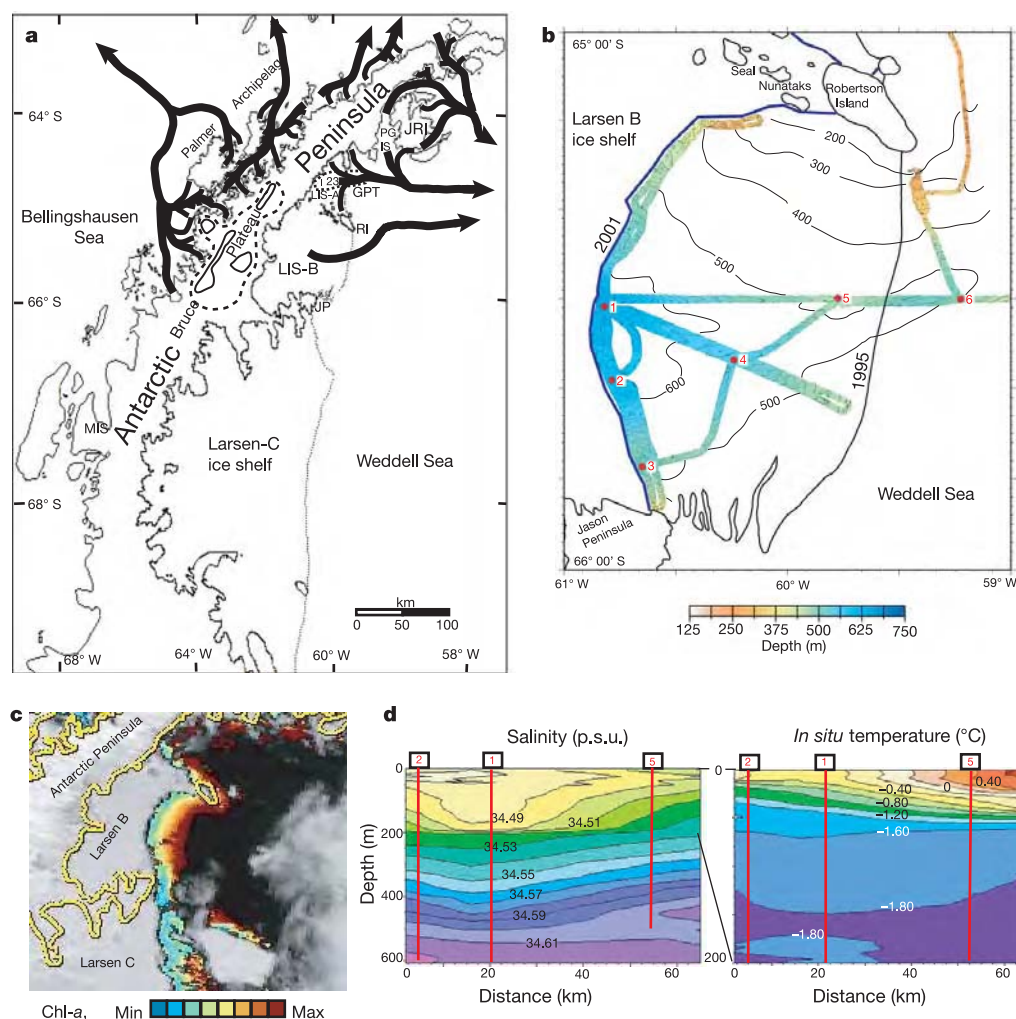


Figure 1 | Location maps, satellite imagery and ocean profiles. **a**, Study area along the northern Antarctic Peninsula, showing reconstructed ice flow lines (arrows), and Larsen A ice shelf (LIS-A), Larsen B ice shelf (LIS-B), Müller ice shelf (MIS), Robertson island (RI), Greenpeace trough (GPT), James Ross island (JRI), Prince Gustav ice shelf (PGIS), and location of core KC-23 (see Fig. 3). Bruce plateau shows proposed (dashed line) extent of high

elevation ice dome and includes modern ice domes of low elevation (solid line). **b**, Bathymetry and core locations (red dots) with positions of LIS-B in 1995 and late 2001 indicated. **c**, Sea surface chlorophyll-*a* (chl-*a*) concentrations from satellite (SeaWiFS) imagery collected on 11 December 2001. **d**, Contoured sections of ocean temperature and salinity from vertical profiles (stations as in **b**).

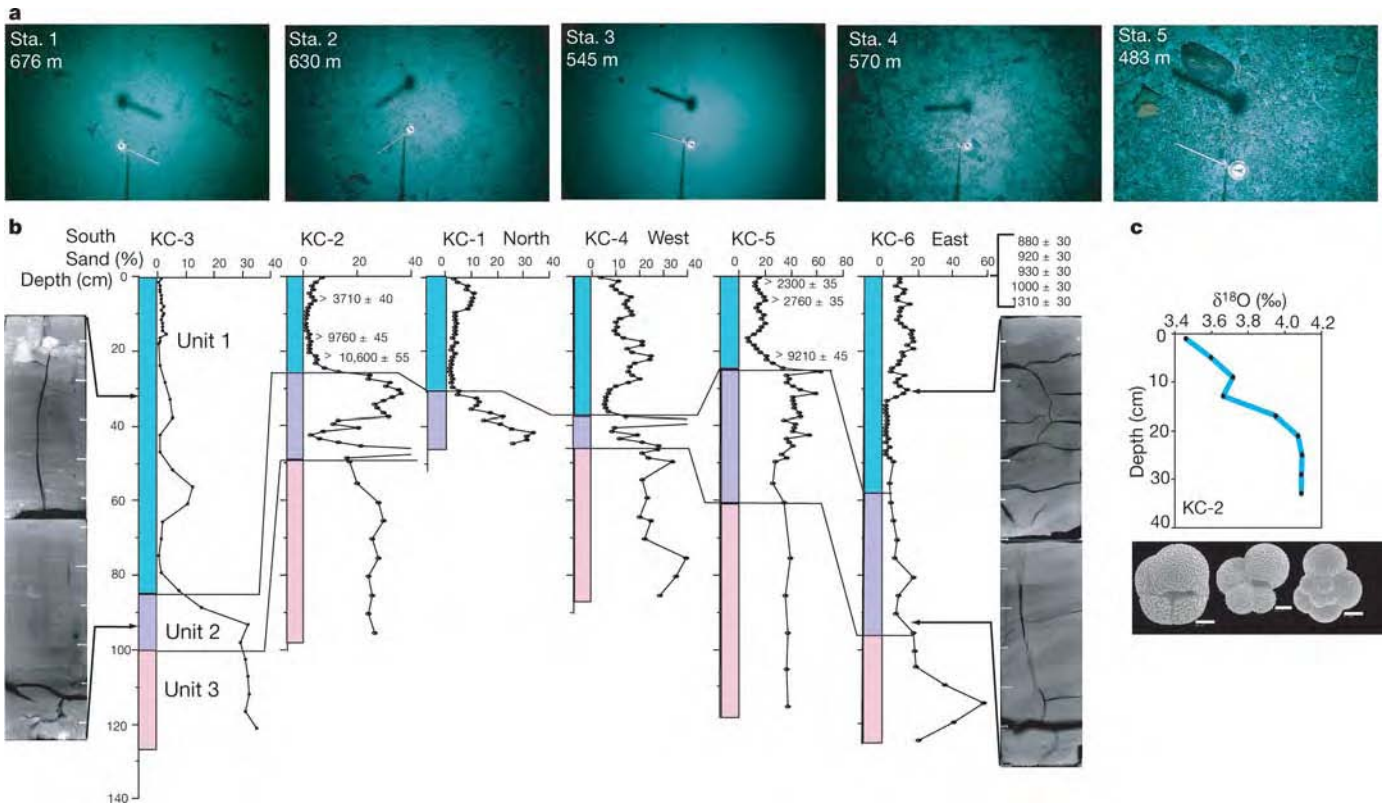


Figure 2 | Bottom photographs and core stratigraphy for sea-floor stations within the Larsen B embayment. **a**, Photographs with current vane and compass for stations 1, 2, 3, 4 and 5. **b**, Stratigraphic fence diagram of kasten cores 1–6 (see Fig. 1b), illustrating sand content, divisions between units 1, 2 and 3, and X-ray radiography negatives for portions of cores 3 and 6.

Uncorrected radiocarbon ages (see Table 1) are shown for biogenic calcite for indicated depths within cores 2 and 5 and for epibenthic organisms at surface of station 6. **c**, Oxygen isotopes on two forms of *Neogloboquadrina pachyderma* (SEM image scale bar is 80 μm for left image, and 60 μm for middle and right image).

history of the LIS-B. A thick LIS-B following ice sheet decoupling could have provided a significant buffer to the water hammer process whereby ice shelves disintegrate¹³. Basal melt rates of 0.78 m yr^{-1} estimated for the LIS-B²² are crucial to this argument, in that LIS-B could not have been sustained through the Holocene without calling on an initial ice shelf that was unrealistically thick or on equally unreasonable flow velocities. It is unlikely that relatively warm ocean waters at any time in the past interacted with the melt regime of the LIS-B. Instead, we propose a slow, steady reduction of ice shelf thickness by a few tens of metres as a response to the natural drawdown of the glacial profile following deglaciation. With continued flow and drainage of ice formed during the Latest Pleistocene

(Fig. 1a), as well as recent under-melting and/or densification²², the LIS-B eventually thinned to the point where it too succumbed to the prolonged period of regional warming now affecting the entire Antarctic Peninsula region^{13,14}. Oxygen isotopes from planktonic foraminifera (Fig. 2c) provide evidence of progressive melt effects within the water column beneath the LIS-B. Over the last 10 kyr a 0.8‰ enrichment of $\delta^{18}\text{O}$ is observed. This is consistent with a combination of ice volume effects for the Holocene (about 0.6‰)^{23,24} and a slight freshening of the water column consistent with limited under-melt. Temperature changes are not responsible for the shift in $\delta^{18}\text{O}$ because the waters today are already close to freezing²⁵ (Fig. 2d) and could not have been

Table 1 | Radiocarbon ages from sediment samples acquired during NBP01-07

Laboratory number	Core depth (cm)	Uncorrected ^{14}C age (yr)	$\delta^{13}\text{C}$ of carbon (‰)	Interval sedimentation rate (mm yr^{-1})	Organic carbon source
AA-49355	G-6 (0)	880 ± 30	0.10		Foraminifera
AA-49357	G-6 (0)	920 ± 30	0.80		Coral
AA-49358	G-6 (0)	930 ± 30	1.50		Bryozoan
AA-49353	G-6 (0)	$1,000 \pm 30$	0.40		Serpulid
AA-49356	G-6 (0)	$1,310 \pm 30$	0.80		Brachiopod
AA-49359	KC-5 (0–2)	$9,360 \pm 100$	-23.40	–	a.i.o.m.
AA-49360	KC-5 (12–14)	$13,930 \pm 120$	-23.60	0.028	a.i.o.m.
AA-49361	KC-5 (24–25)	$18,680 \pm 170$	-23.70	0.021	a.i.o.m.
NO-38246	KC-5 (0–2)	$2,300 \pm 35$	1.03	0.007	Foraminifera
NO-39000	KC-5 (4–6)	$2,760 \pm 35$	0.91	0.088	Foraminifera
NO-39001	KC-5 (20–22)	$9,210 \pm 45$	0.92	0.025	Foraminifera
NO-38996	KC-2 (4–6)	$3,710 \pm 40$	0.56	0.018	Foraminifera
NO-38244	KC-2 (16–18)	$9,760 \pm 45$	0.35	0.020	Foraminifera
NO-38245	KC-2 (20–22)	$10,600 \pm 55$	-0.05	0.048	Foraminifera

a.i.o.m., acid insoluble organic matter.

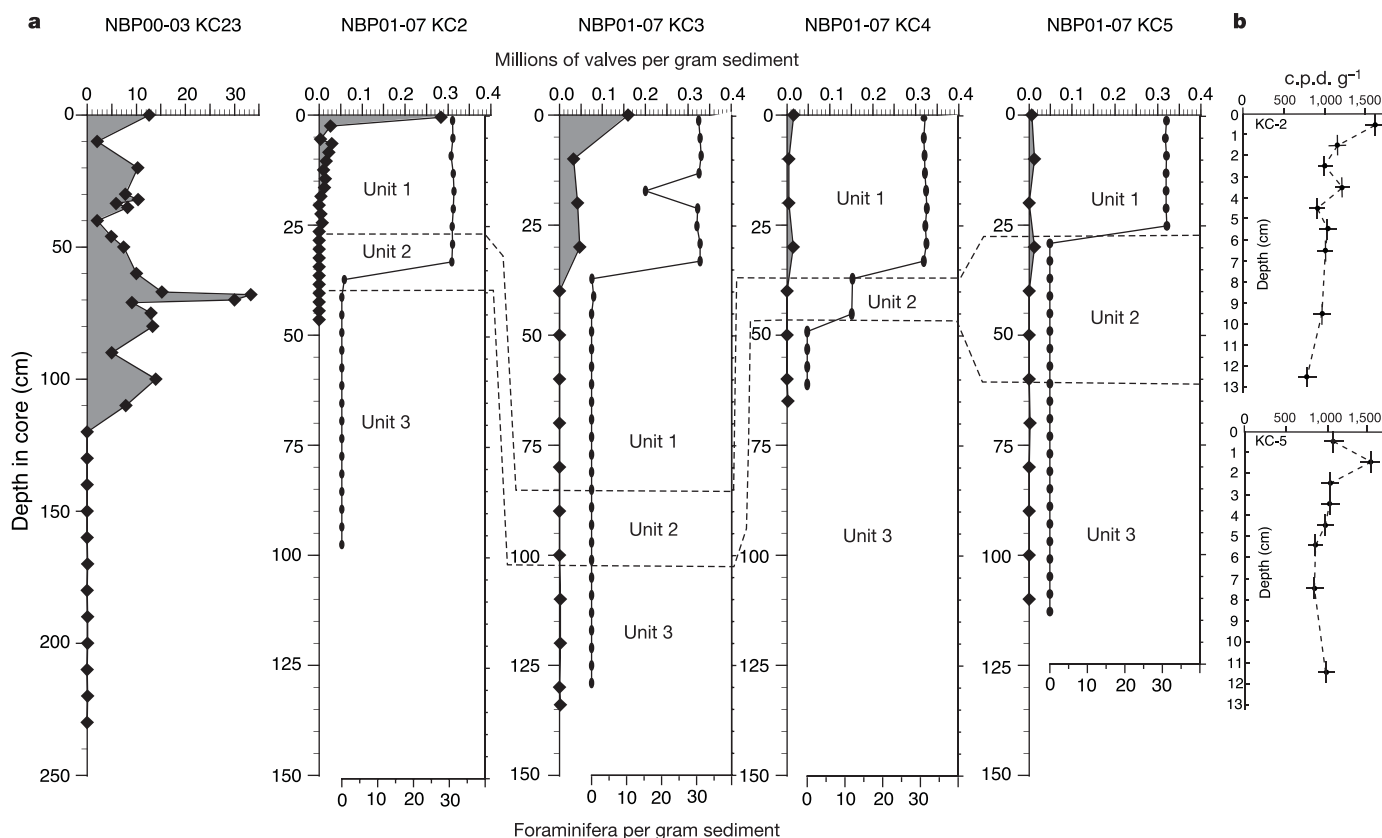


Figure 3 | Microfossil abundance and radioactive ^{210}Pb activity within marine sediment cores from beneath the Larsen ice shelf. **a**, Stratigraphic abundance of diatom valves (top scale, diamonds) and foraminifera (bottom scale, ellipses) with stratigraphic unit divisions (see Figs 1b and 2b). **b**, ^{210}Pb

activities for near surface samples in cores 2 and 5 (units on x axis for ^{210}Pb data are counts per day per gram (c.p.d. g^{-1}). Error bars for the ^{210}Pb data represent the sample depths (1 cm) and counting statistics based on a Poisson distribution and incremental background measurements.

significantly colder during the Holocene. Oceanographic data collected along the front of the LIS-B in December 2001 show a lens of sub-ice shelf melt water consistent with, at most, a few centimetres of under-melt beneath the ice shelf²⁶ (Fig. 1d).

Direct geologic evidence from beneath the LIS-B indicates that the recent break-up event is unprecedented in the Holocene history of this glacial system. The ice shelf's demise is probably the consequence of a combination of long term thinning (by a few tens of metres) over thousands of years and short term (multi-decadal) cumulative increases in surface air temperature that have exceeded the natural variation of regional climate during the Holocene interglacial.

METHODS

Field methods. Cores, water samples, bottom photographs and oceanographic data were collected during US Antarctic Program cruise 2001-07 of the *Nathaniel B. Palmer* (see <http://www.hamilton.edu/news/exp/Antarctica2001/> for details).

Diatom analyses. Quantitative diatom slides were prepared using a settling method²⁷ that produces a random distribution of diatoms and allows calculation of absolute diatom abundances. Diatoms were counted along transects at $400\times$ magnification on an Olympus BX60 microscope. Between 5 and 10 transects were counted for each sample.

Radioisotope dating. The chronology of radiocarbon dates in Antarctica depends upon an assumption of a large reservoir correction (~ 1.2 kyr) and adjustments for reworking of organic matter¹⁶. We avoid the later problem by using calcareous foraminifera (Fig. 2c), listing all ages as uncorrected for reservoir age, and providing living ages (reservoir age estimate) on modern calcite from the locale (Table 1). The ages reported in Table 1 are corrected only for the $\delta^{13}\text{C}$ fractionation and are not calibrated. Processing of samples follows established procedures^{15,16}. Interval sedimentation rates are calculated from ages between adjacent stratigraphic levels and are provided for both calcite and acid insoluble organic matter (a.i.o.m.) samples. ^{14}C analyses were completed at the

University of Arizona (AA) and the National Ocean Sciences Accelerator Mass Spectrometer facility at Woods Hole Oceanographic Institution (NO). ^{210}Pb methods followed published procedures²⁸.

Manganese oxide. Eleven thickness measurements of MnO_2 , ranging from 4.2 to $16.7\ \mu\text{m}$, were determined from thin sections of clasts collected at site 5. We used a FEI Quanta scanning electron microscope integrated with an EDAX Phoenix energy dispersive spectrometer. The modal thickness of MnO_2 coating was $3\text{--}6\ \mu\text{m}$ with a mean value of $9.4\ \mu\text{m}$. We then compared the mean thicknesses to published rates (and means) of MnO_2 accretion near the Weddell Sea¹⁸ to provide an estimate for the duration of exposure as between 1.5 and 2.7 kyr.

Grain lithology. Supplementary Data illustrate downcore petrographic assignment of the gravel to coarse sand fraction from sample intervals in core 5 (Fig. 2b). Gravel and coarse sand fractions were removed from matrix by wet sieving, and grain counts were made by examination under a binocular microscope.

Glacial reconstruction. Ice sheet flow directions and reconstructions (Fig. 1c) were based upon a compilation of published information on swath bathymetry for regions surrounding the northern Antarctic Peninsula (Supplementary Information). Details of sea-floor relief for the LIS-B region are confirmed by comparisons of our data (Fig. 1b) to published information²⁹.

Received 16 December 2004; accepted 7 June 2005.

1. Mercer, J. H. West Antarctic ice sheet and CO_2 greenhouse effect: a threat of disaster. *Nature* **271**, 321–325 (1978).
2. Vaughan, D. G. & Doake, C. S. M. Recent atmospheric warming and retreat of ice shelves on the Antarctic Peninsula. *Nature* **379**, 328–331 (1996).
3. Scambos, T., Hulbe, C. & Fahnestock, M. Climate-induced ice shelf disintegration in the Antarctic Peninsula. *Antarct. Res. Ser.* **79**, 79–92 (2003).
4. De Angelis, H. & Skvarca, P. Glacier surge after ice shelf collapse. *Science* **299**, 1560–1562 (2003).
5. Scambos, T. A., Bohlander, J. A., Shuman, C. A. & Skvarca, P. Glacier acceleration and thinning after ice shelf collapse in the Larsen B embayment, Antarctica. *Geophys. Res. Lett.* **31**, L18402, doi:10.1029/2004GL020670 (2004).

6. Rignot, E. *et al.* Accelerated ice discharge from the Antarctic Peninsula following the collapse of Larsen B ice shelf. *Geophys. Res. Lett.* **31**, L18401, doi:10.1029/2004GL020697 (2004).
7. Clapperton, C. M. & Sugden, D. E. Late Quaternary glacial history of the George VI Sound area, West Antarctica. *Quat. Res.* **18**, 243–267 (1982).
8. Domack, E. W. *et al.* Late Holocene advance of the Müller Ice Shelf, Antarctic Peninsula: sedimentologic, geochemical, and palaeontological evidence. *Antarct. Sci.* **7**, 159–170 (1995).
9. Hjort, C., Bentley, M. J. & Ingólfsson, O. Holocene and pre-Holocene temporary disappearance of the George VI Ice Shelf, Antarctic Peninsula. *Antarct. Sci.* **13**, 296–301 (2001).
10. Pudsey, C. J. & Evans, J. First survey of Antarctic sub-ice shelf sediments reveals mid-Holocene ice shelf retreat. *Geol.* **29**, 787–790 (2001).
11. Brachfeld, S. *et al.* Holocene history of the Larsen-A Ice Shelf constrained by geomagnetic paleointensity dating. *Geology* **31**, 749–752 (2003).
12. Morris, E. M. & Vaughan, D. G. Glaciologic climate relationships spatial and temporal variation of surface temperature on the Antarctic Peninsula and the limit of viability of ice shelves. *Antarct. Res. Ser.* **79**, 61–68 (2003).
13. Scambos, T. A., Hulbe, C., Fahnestock, M. & Bohlander, J. The link between climate warming and break-up of ice shelves in the Antarctic Peninsula. *J. Glaciol.* **154**, 516–530 (2001).
14. King, J. *et al.* Antarctic Peninsula climate variability and its causes as revealed by analysis of instrumental records. *Antarct. Res. Ser.* **79**, 17–30 (2003).
15. Gilbert, R. & Domack, E. W. The sedimentary record of disintegrating ice shelves in a warming climate, Antarctic Peninsula. *Geochem. Geophys. Geosyst.* **4**, 1038, doi:10.1029/2002GC000441 (2003).
16. Domack, E. W., Jacobson, E. A., Shipp, S. S. & Anderson, J. B. Late Pleistocene-Holocene retreat of the West Antarctic Ice Sheet system in the Ross Sea: Part 2 — sedimentologic and stratigraphic signature. *Geol. Soc. Am. Bull.* **111**, 1517–1536 (1999).
17. MacAyeal, D. R., Scambos, T. A., Hulbe, C. & Fahnestock, M. A. Catastrophic ice-shelf break-up by an ice-shelf-fragment capsize mechanism. *J. Glaciol.* **49**, 22–36 (2003).
18. Frank, M. *et al.* North Atlantic Deep Water export to the Southern Ocean over the past 14 Myr: evidence from Nd and Pb isotopes in ferromanganese crusts. *Paleoceanography* **17**, doi: 10.1029/2000PA000606 (2002).
19. Ishman, S. E. & Szymczek, P. Foraminiferal distributions in the former Larsen-A ice shelf and Prince Gustav Channel region, eastern Antarctic Peninsula margin: a baseline for Holocene paleoenvironmental change. *Antarct. Res. Ser.* **79**, 239–260 (2003).
20. Jacobs, S. S., Gordon, A. L. & Ardai, J. L. Jr. Circulation and melting beneath the Ross Ice Shelf. *Science* **203**, 439–445 (1979).
21. Gilbert, R., Domack, E. W. & Camerlenghi, A. Deglacial history of the Greenpeace Trough: ice sheet to ice shelf transition in the northwestern Weddell Sea. *Antarct. Res. Ser.* **79**, 195–204 (2003).
22. Shepherd, A., Wingham, D., Payne, T. & Skvarca, P. Larsen Ice Shelf has progressively thinned. *Science* **302**, 856–859 (2003).
23. Pahnke, K., Zahn, R., Elderfield, H. & Schultz, M. 340,000-year centennial-scale marine record of Southern Hemisphere climate oscillation. *Science* **301**, 948–952 (2003).
24. Norris, R. D., Park, B. K., Kang, S. H. & Khim, B. K. Stable isotope and ecological habitat of planktonic foraminifera adjacent to the ice edge in the western Weddell Sea. *Geosci. J.* **2**, 88–98 (1998).
25. Nicholls, K. W. *et al.* Water mass modification over the continental shelf north of Ronne Ice Shelf, Antarctica. *J. Geophys. Res.* **108**, doi:10.1029/2002JC001713 (2003).
26. Doane, S. S. *Oceanographic Observations in Front of the Larsen B Ice Shelf; Antarctica Prior to its Collapse in 2002*. BA thesis, Hamilton College (2003).
27. Scherer, R. P. A new method for the determination of absolute abundance of diatoms and other silt-sized sedimentary particles. *J. Paleolimnol.* **12**, 171–179 (1994).
28. Domack, E. W. *et al.* Marine sedimentary record of natural variability and recent warming in the Antarctic Peninsula. *Antarct. Res. Ser.* **79**, 205–224 (2003).
29. Evans, J. *et al.* Late Quaternary glacial history, flow dynamics and sedimentation along the eastern margin of the Antarctic Peninsula Ice Sheet. *Quat. Sci. Rev.* **24**, 741–774 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by grants from the National Science Foundation Office of Polar Programs. Timely support of analytical needs at the University of Arizona Accelerator Facility and the National Ocean Sciences Accelerator Mass Spectrometer laboratory was appreciated, as were the contributions of H. Schrum, E. Backman and K. Bart, and the comments by M. Canals, E. Rignot, L. Padman and T. Scambos.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to E.D. (edomack@hamilton.edu).

LETTERS

Increasing destructiveness of tropical cyclones over the past 30 years

Kerry Emanuel¹

Theory¹ and modelling² predict that hurricane intensity should increase with increasing global mean temperatures, but work on the detection of trends in hurricane activity has focused mostly on their frequency^{3,4} and shows no trend. Here I define an index of the potential destructiveness of hurricanes based on the total dissipation of power, integrated over the lifetime of the cyclone, and show that this index has increased markedly since the mid-1970s. This trend is due to both longer storm lifetimes and greater storm intensities. I find that the record of net hurricane power dissipation is highly correlated with tropical sea surface temperature, reflecting well-documented climate signals, including multi-decadal oscillations in the North Atlantic and North Pacific, and global warming. My results suggest that future warming may lead to an upward trend in tropical cyclone destructive potential, and—taking into account an increasing coastal population—a substantial increase in hurricane-related losses in the twenty-first century.

Fluctuations in tropical cyclone activity are of obvious importance to society, especially as populations of afflicted areas increase⁵. Tropical cyclones account for a significant fraction of damage, injury and loss of life from natural hazards and are the costliest natural catastrophes in the US⁶. In addition, recent work suggests that global tropical cyclone activity may play an important role in driving the oceans' thermohaline circulation, which has an important influence on regional and global climate⁷.

Studies of tropical cyclone variability in the North Atlantic reveal large interannual and interdecadal swings in storm frequency that have been linked to such regional climate phenomena as the El Niño/Southern Oscillation⁸, the stratospheric quasi-biennial oscillation⁹, and multi-decadal oscillations in the North Atlantic region¹⁰. Variability in other ocean basins is less well documented, perhaps because the historical record is less complete.

Concerns about the possible effects of global warming on tropical cyclone activity have motivated a number of theoretical, modelling and empirical studies. Basic theory¹¹ establishes a quantitative upper bound on hurricane intensity, as measured by maximum surface wind speed, and empirical studies show that when accumulated over large enough samples, the statistics of hurricane intensity are strongly controlled by this theoretical potential intensity¹². Global climate models show a substantial increase in potential intensity with anthropogenic global warming, leading to the prediction that actual storm intensity should increase with time¹. This prediction has been echoed in climate change assessments¹³. A recent comprehensive study using a detailed numerical hurricane model run using climate predictions from a variety of different global climate models² supports the theoretical predictions regarding changes in storm intensity. With the observed warming of the tropics of around 0.5 °C, however, the predicted changes are too small to have been observed, given limitations on tropical cyclone intensity estimation.

The issue of climatic control of tropical storm frequency is far

more controversial, with little guidance from existing theory. Global climate model predictions of the influence of global warming on storm frequency are highly inconsistent, and there is no detectable trend in the global annual frequency of tropical cyclones in historical tropical cyclone data.

Although the frequency of tropical cyclones is an important scientific issue, it is not by itself an optimal measure of tropical cyclone threat. The actual monetary loss in wind storms rises roughly as the cube of the wind speed¹⁴ as does the total power dissipation (PD; ref. 15), which, integrated over the surface area affected by a storm and over its lifetime is given by:

$$PD = 2\pi \int_0^{\tau} \int_0^{r_0} C_D \rho |V|^3 r dr dt \quad (1)$$

where C_D is the surface drag coefficient, ρ is the surface air density, $|V|$ is the magnitude of the surface wind, and the integral is over radius to an outer storm limit given by r_0 and over τ , the lifetime of the storm. The quantity PD has the units of energy and reflects the total power dissipated by a storm over its life. Unfortunately, the area integral in equation (1) is difficult to evaluate using historical data sets, which seldom report storm dimensions. On the other hand, detailed studies show that radial profiles of wind speed are generally geometrically similar¹⁶ whereas the peak wind speeds exhibit little if any correlation with measures of storm dimensions¹⁷. Thus variations in storm size would appear to introduce random errors in an evaluation of equation (1) that assumes fixed storm dimensions. In the integrand of equation (1), the surface air density varies over roughly 15%, while the drag coefficient is thought to increase over roughly a factor of two with wind speed, but levelling off at wind speeds in excess of about 30 m s⁻¹ (ref. 18). As the integral in equation (1) will, in practice, be dominated by high wind speeds, we approximate the product $C_D \rho$ as a constant and define a simplified power dissipation index as:

$$PDI \equiv \int_0^{\tau} V_{\max}^3 dt \quad (2)$$

where $V_{\max}$ is the maximum sustained wind speed at the conventional measurement altitude of 10 m. Although not a perfect measure of net power dissipation, this index is a better indicator of tropical cyclone threat than storm frequency or intensity alone. Also, the total power dissipation is of direct interest from the point of view of tropical cyclone contributions to upper ocean mixing and the thermohaline circulation⁷. This index is similar to the 'accumulated cyclone energy' (ACE) index¹⁹, defined as the sum of the squares of the maximum wind speed over the period containing hurricane-force winds.

The analysis technique, data sources, and corrections to the raw data are described in the Methods section and in Supplementary Methods. To emphasize long-term trends and interdecadal variability, the PDI is accumulated over an entire year and, individually, over

¹Program in Atmospheres, Oceans, and Climate, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.

each of several major cyclone-prone regions. To minimize the effect of interannual variability, we apply to the time series of annual PDI a 1-2-1 smoother defined by:

$$x'_i = 0.25(x_{i-1} + x_{i+1}) + 0.5x_i \quad (3)$$

where x_i is the value of the variable in year i and x'_i is the smoothed value. This filter is generally applied twice in succession.

Figure 1 shows the PDI for the North Atlantic and the September mean tropical sea surface temperature (SST) averaged over one of the prime genesis regions in the North Atlantic²⁰. There is an obvious strong relationship between the two time series ($r^2 = 0.65$), suggesting that tropical SST exerts a strong control on the power dissipation index. The Atlantic multi-decadal mode discussed in ref. 10 is evident in the SST series, as well as shorter period oscillations possibly related to the El Niño/Southern Oscillation and the North Atlantic Oscillation. But the large upswing in the last decade is unprecedented, and probably reflects the effect of global warming. We will return to this subject below.

Figure 2 shows the annually accumulated, smoothed PDI for the western North Pacific, together with July–November average smoothed SST in a primary genesis region for the North Pacific. As in the Atlantic, these are strongly correlated, with an r^2 of 0.63. Some of the interdecadal variability is associated with the El Niño/Southern Oscillation, as documented by Camargo and Sobel¹⁹. The SST time series shows that the upswing in SST since around 1975 is unusual by the standard of the past 70 yr.

There are reasons to believe that global tropical SST trends may have less effect on tropical cyclones than regional fluctuations, as tropical cyclone potential intensity is sensitive to the difference between SST and average tropospheric temperature. In an effort to quantify a global signal, annual average smoothed SST between 30° N and 30° S is compared to the sum of the North Atlantic and western North Pacific smoothed PDI values in Fig. 3. The two time series are correlated with an r^2 of 0.69. The upturn in tropical mean surface temperature since 1975 has been generally ascribed to global warming, suggesting that the upward trend in tropical cyclone PDI values is at least partially anthropogenic. It is interesting that this trend has involved more than a doubling of North Atlantic plus western North Pacific PDI over the past 30 yr.

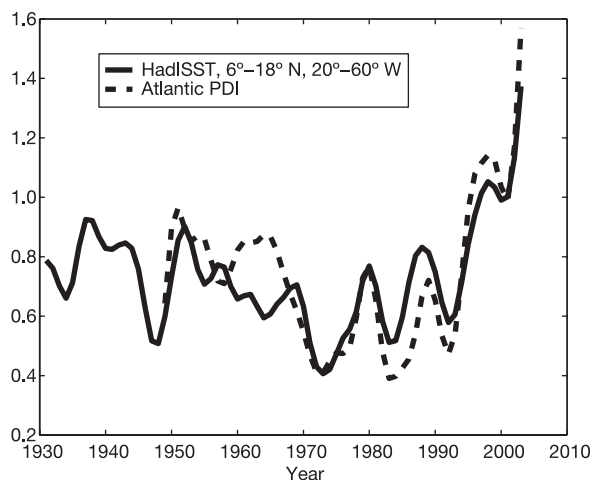


Figure 1 | A measure of the total power dissipated annually by tropical cyclones in the North Atlantic (the power dissipation index, PDI) compared to September sea surface temperature (SST). The PDI has been multiplied by 2.1×10^{-12} and the SST, obtained from the Hadley Centre Sea Ice and SST data set (HadISST)²², is averaged over a box bounded in latitude by 6° N and 18° N, and in longitude by 20° W and 60° W. Both quantities have been smoothed twice using equation (3), and a constant offset has been added to the temperature data for ease of comparison. Note that total Atlantic hurricane power dissipation has more than doubled in the past 30 yr.

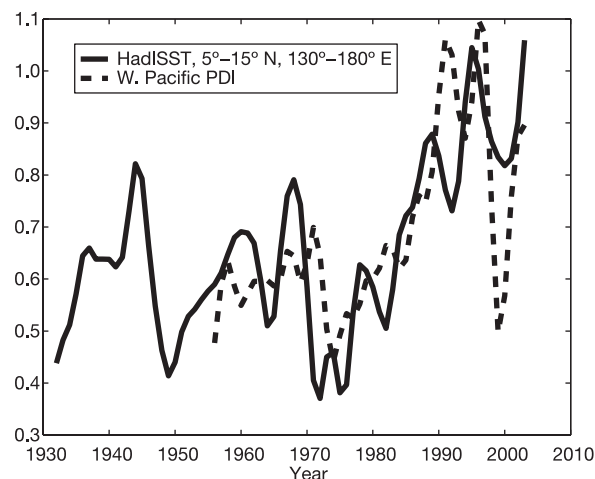


Figure 2 | Annually accumulated PDI for the western North Pacific, compared to July–November average SST. The PDI has been multiplied by a factor of 8.3×10^{-13} and the HadISST (with a constant offset) is averaged over a box bounded in latitude by 5° N and 15° N, and in longitude by 130° E and 180° E. Both quantities have been smoothed twice using equation (3). Power dissipation by western North Pacific tropical cyclones has increased by about 75% in the past 30 yr.

The large increase in power dissipation over the past 30 yr or so may be because storms have become more intense, on the average, and/or have survived at high intensity for longer periods of time. The accumulated annual duration of storms in the North Atlantic and western North Pacific has indeed increased by roughly 60% since 1949, though this may partially reflect changes in reporting practices, as discussed in Methods. The annual average storm peak wind speed summed over the North Atlantic and eastern and western North Pacific has also increased during this period, by about 50%. Thus both duration and peak intensity trends are contributing to the overall increase in net power dissipation. For fixed rates of intensification and dissipation, storms will take longer to reach greater peak winds, and also take longer to dissipate. Thus, not surprisingly, stronger storms last longer; times series of duration and peak intensity are correlated with an r^2 of 0.74.

In theory, the peak wind speed of tropical cyclones should increase

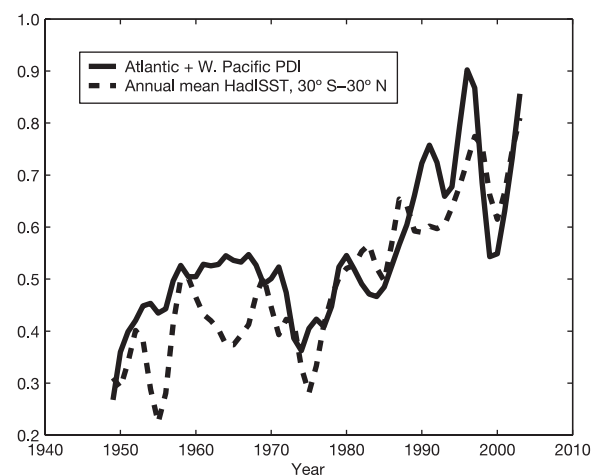


Figure 3 | Annually accumulated PDI for the western North Pacific and North Atlantic, compared to annually averaged SST. The PDI has been multiplied by a factor of 5.8×10^{-13} and the HadISST (with a constant offset) is averaged between 30° S and 30° N. Both quantities have been smoothed twice using equation (3). This combined PDI has nearly doubled over the past 30 yr.

by about 5% for every 1 °C increase in tropical ocean temperature¹. Given that the observed increase has only been about 0.5 °C, these peak winds should have only increased by 2–3%, and the power dissipation therefore by 6–9%. When coupled with the expected increase in storm lifetime, one might expect a total increase of PDI of around 8–12%, far short of the observed change.

Tropical cyclones do not respond directly to SST, however, and the appropriate measure of their thermodynamic environment is the potential intensity, which depends not only on surface temperature but on the whole temperature profile of the troposphere. I used daily averaged re-analysis data and Hadley Centre SST to re-construct the potential maximum wind speed, and then averaged the result over each calendar year and over the same tropical areas used to calculate the average SST. In both the Atlantic and western North Pacific, the time series of potential intensity closely follows the SST, but increases by about 10% over the period of record, rather than the predicted 2–3%. Close examination of the re-analysis data shows that the observed atmospheric temperature does not keep pace with SST. This has the effect of increasing the potential intensity. Given the observed increase of about 10%, the expected increase of PDI is about 40%, taking into account the increased duration of events. This is still short of the observed increase.

The above discussion suggests that only part of the observed increase in tropical cyclone power dissipation is directly due to increased SSTs; the rest can only be explained by changes in other factors known to influence hurricane intensity, such as vertical wind shear. Analysis of the 250–850 hPa wind shear from reanalysis data, over the same portion of the North Atlantic used to construct Fig. 1, indeed shows a downward trend of 0.3 m s⁻¹ per decade over the period 1949–2003, but most of this decrease occurred before 1970, and at any rate the decrease is too small to have had much effect. Tropical cyclone intensity also depends on the temperature distribution of the upper ocean, and there is some indication that sub-surface temperatures have also been increasing²¹, thereby reducing the negative feedback from storm-induced mixing.

Whatever the cause, the near doubling of power dissipation over the period of record should be a matter of some concern, as it is a measure of the destructive potential of tropical cyclones. Moreover, if upper ocean mixing by tropical cyclones is an important contributor to the thermohaline circulation, as hypothesized by the author⁷, then global warming should result in an increase in the circulation and therefore an increase in oceanic enthalpy transport from the tropics to higher latitudes.

METHODS

Positions and maximum sustained surface winds of tropical cyclones are reported every six hours as part of the 'best track' tropical data sets. (In the data sets used here, from the US Navy's Joint Typhoon Warning Center (JTWC) and the National Oceanographic and Atmospheric Administration's National Hurricane Center (NHC), 'maximum sustained wind' is defined as the one-minute average wind speed at an altitude of 10 m.) For the Atlantic, and eastern and central North Pacific, these data are available from the NHC, while for the western North Pacific, the northern Indian Ocean, and all of the Southern Hemisphere, data from JTWC were used.

Owing to changes in measuring and reporting practices since systematic observations of tropical cyclones began in the mid-1940s, there are systematic biases in reported tropical cyclone wind speeds that must be accounted for in

analysing trends. The sources of these biases and corrections made to account for them are described in Supplementary Methods.

Received 28 January; accepted 3 June 2005.

Published online 31 July 2005.

1. Emanuel, K. A. The dependence of hurricane intensity on climate. *Nature* **326**, 483–485 (1987).
2. Knutson, T. R. & Tuleya, R. E. Impact of CO₂-induced warming on simulated hurricane intensity and precipitation: Sensitivity to the choice of climate model and convective parameterization. *J. Clim.* **17**, 3477–3495 (2004).
3. Landsea, C. W., Nicholls, N., Gray, W. M. & Avila, L. A. Downward trends in the frequency of intense Atlantic hurricanes during the past five decades. *Geophys. Res. Lett.* **23**, 1697–1700 (1996).
4. Chan, J. C. L. & Shi, J.-E. Long-term trends and interannual variability in tropical cyclone activity over the western North Pacific. *Geophys. Res. Lett.* **23**, 2765–2767 (1996).
5. Pielke, R. A. J., Rubiera, J., Landsea, C. W., Fernandez, M. L. & Klein, R. Hurricane vulnerability in Latin America and the Caribbean: Normalized damage and loss potentials. *Nat. Hazards Rev.* **4**, 101–114 (2003).
6. Pielke, R. A. J. & Landsea, C. W. Normalized U.S. hurricane damage, 1925–1995. *Weath. Forecast.* **13**, 621–631 (1998).
7. Emanuel, K. A. The contribution of tropical cyclones to the oceans' meridional heat transport. *J. Geophys. Res.* **106**, 14771–14782 (2001).
8. Pielke, R. A. J. & Landsea, C. W. La Niña, El Niño, and Atlantic hurricane damages in the United States. *Bull. Am. Meteorol. Soc.* **80**, 2027–2033 (1999).
9. Gray, W. M. Atlantic seasonal hurricane frequency. Part I: El Niño and 30 mb quasi-biennial oscillation influences. *Mon. Weath. Rev.* **112**, 1649–1668 (1984).
10. Goldenberg, S. B., Landsea, C. W., Mestas-Núñez, A. M. & Gray, W. M. The recent increase in Atlantic hurricane activity: Causes and implications. *Science* **293**, 474–479 (2001).
11. Bister, M. & Emanuel, K. A. Dissipative heating and hurricane intensity. *Meteorol. Atmos. Phys.* **50**, 233–240 (1998).
12. Emanuel, K. A. A statistical analysis of tropical cyclone intensity. *Mon. Weath. Rev.* **128**, 1139–1152 (2000).
13. Henderson-Sellers, A. *et al.* Tropical cyclones and global climate change: A post-IPCC assessment. *Bull. Am. Meteorol. Soc.* **79**, 19–38 (1998).
14. Southern, R. L. The global socio-economic impact of tropical cyclones. *Aust. Meteorol. Mag.* **27**, 175–195 (1979).
15. Emanuel, K. A. The power of a hurricane: An example of reckless driving on the information superhighway. *Weather* **54**, 107–108 (1998).
16. Mallen, K. J., Montgomery, M. T. & Wang, B. Re-examining the near-core radial structure of the tropical cyclone primary circulation: Implications for vortex resiliency. *J. Atmos. Sci.* **62**, 408–425 (2005).
17. Weatherford, C. L. & Gray, W. M. Typhoon structure as revealed by aircraft reconnaissance. Part I: Data analysis and climatology. *Mon. Weath. Rev.* **116**, 1032–1043 (1988).
18. Powell, M. D., Vickery, P. J. & Reinhold, T. A. Reduced drag coefficients for high wind speeds in tropical cyclones. *Nature* **422**, 279–283 (2003).
19. Camargo, S. J. & Sobel, A. H. Western North Pacific tropical cyclone intensity and ENSO. *J. Clim.* (in the press).
20. Saunders, M. A. & Harris, A. R. Statistical evidence links exceptional 1995 Atlantic hurricane season to record sea warming. *Geophys. Res. Lett.* **24**, 1255–1258 (1997).
21. Levitus, S., Antonov, J. I., Boyer, T. P. & Stephens, C. Warming of the world ocean. *Science* **287**, 2225–2229 (2000).
22. Rayner, N. A. *et al.* Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century. *J. Geophys. Res.* **108**, 4407, doi:10.1029/2002JD002670 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The author is grateful for correspondence with S. Camargo, C. Guard, C. Landsea and A. Sobel.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to the author at emanuel@texmex.mit.edu.

Earthquake slip weakening and asperities explained by thermal pressurization

Christopher A. J. Wibberley¹ & Toshihiko Shimamoto²

An earthquake occurs when a fault weakens during the early portion of its slip at a faster rate than the release of tectonic stress driving the fault motion^{1,2}. This slip weakening occurs over a critical distance, D_c . Understanding the controls on D_c in nature is severely limited, however, because the physical mechanism of weakening is unconstrained. Conventional friction experiments, typically conducted at slow slip rates and small displacements, have obtained D_c values that are orders of magnitude lower than values estimated from modelling seismological data for natural earthquakes^{3–6}. Here we present data on fluid transport properties of slip zone rocks and on the slip zone width in the centre of the Median Tectonic Line fault zone, Japan. We show that the discrepancy between laboratory and seismological results can be resolved if thermal pressurization of the pore fluid^{7–10} is the slip-weakening mechanism. Our analysis indicates that a planar fault segment with an impermeable and narrow slip zone will become very unstable during slip and is likely to be the site of a seismic asperity.

A simple model leading to unstable fault motion is the so-called slip-weakening model¹¹, in which a fault linearly weakens over a critical distance, D_c . Seismological studies in the past decade suggest that D_c is around 0.05–1.0 m (refs 4–6), whereas experiments at low slip rate yield D_c values typically of tens of micrometres^{3,12}. The critical distance parameter in the rate-and-state frictional constitutive law¹³ is also orders of magnitude smaller than seismically determined D_c . Previous attempts to explain this large difference in D_c have concentrated only on scale effects, such as the topography of the sliding surfaces^{12,14} and slip zone width¹⁵, rather than the mechanism for slip weakening. Yet frictional heating is inferred from some exhumed faults by the presence of pseudotachylytes¹⁶ and changes in electron spin resonance signals in fault gouge¹⁷. Furthermore, slip zones are now recognized usually to consist of granular material that is likely to be saturated in fluid^{18,19}. Thus, one possible mechanism for dramatic fault weakening is the thermal pressurization process^{7–10}, in that frictional heating generates excess fluid pressures and lowers effective shear strength. However, the poor constraints on widths and fluid transport properties of earthquake slip zones have made it impossible to include this process in the analyses of earthquake rupture propagation.

We investigated the structure and fluid transport properties of a continuous slip zone in a well-exposed exhumed fault zone, the Median Tectonic Line (MTL) in Mie prefecture, Japan^{18,20}. A consistent feature with active seismogenic faults in Japan, and exhumed fault zones elsewhere¹⁹, is the presence of a sharp, continuous, planar central slip zone of highly sheared clay-rich gouge thought to have formed during the latest seismogenic slip events on the fault (Fig. 1). The continuity of the central slip zone in the outcrop observed is very striking (Fig. 1b), and it cross-cuts earlier-formed fault rocks and

fault zone structures (Fig. 1c). The slip zone width typically ranges from 10 to 50 mm (Fig. 1d). Similar recent studies of continuous central slip zones in well-known active fault zones in Japan revealed widths of 10 ± 6.1 mm (number of measurements, $n = 28$), 8.6 ± 2.8 mm ($n = 11$) and 15 ± 10 mm ($n = 21$) for three outcrops of the Neodani fault²¹, 5.1 ± 2.5 mm ($n = 28$) for the Iida–Matsukawa fault, and 8.3 ± 6.2 mm ($n = 23$) for the Kan-nawa fault. The slip zone width, the most important scale effect in our thermal pressurization analysis below, is the concentrated zone of shear deformation during an earthquake. This zone is much narrower than the entire zone of incohesive and cohesive fault rocks (typically a few to several hundred metres thick), but is much wider than concentrated deformation zones produced in laboratory friction experiments (typically a few to 100 micrometres)^{15,22}.

The permeability and porosity of clay-rich slip zone gouge samples were measured in the laboratory using nitrogen gas as a pore fluid at 50 MPa, for different effective pressures P_e (where $P_e = \text{lithostatic pressure } (P_{\text{lith}}) - \text{fluid pressure } (P_f)$). Steps of increasing P_e were intended to allow measurements at different simulated depths in the Earth's crust up to about 6 km, assuming hydrostatic fluid pressures, with steps of decreasing P_e simulating rising fluid pressure at that depth. Results show that both permeability (k) and porosity (ϕ) are low compared to other incohesive granular materials (Fig. 2). The permeability data (Fig. 2a) show an anisotropy typical of foliated fault rocks²³, being much lower when perpendicular to the fault and gouge fabric at high P_e than parallel to it. From the best-fit functions of porosity (Fig. 2b), and a best-fit function for bulk compressibility β_{P_e} of $\beta = 2.5 \times 10^{-10} e^{-1.38 \times 10^{-8} P_e}$ obtained in an earlier study²⁴, the storage capacity (β_c) of the fault rock (Fig. 2b) is calculated as a function of P_e , P_f and temperature T :

$$\beta_{c(P_e, P_f, T)} = (\beta_{P_e} - \beta_m) + \phi_{P_e} (\beta_{w(P_f, T)} - \beta_m) \quad (1)$$

where $\beta_{w(P_f, T)}$ is the compressibility of water and β_m is the compressibility of individual grains, assumed to be constant at $1.2 \times 10^{-11} \text{ Pa}^{-1}$. The hydraulic diffusivity (D_h) can then be found as:

$$D_{h(P_e, P_f, T)} = \frac{k_{P_e}}{\eta_{P_f, T} \cdot \beta_{c(P_e, P_f, T)}} \quad (2)$$

where η is the fluid viscosity. The final parameter constrained by our laboratory data is the thermal pressurization coefficient, Γ , being the ratio of fluid volume thermal expansivity to storage capacity for a fixed mass of fluid:

$$\Gamma_{P_e, P_f, T} = \frac{\phi_{P_e} \alpha_w + (1 - \phi_{P_e}) \alpha_m - \alpha_{sf}}{\beta_{c(P_e, P_f, T)}} \quad (3)$$

where α_w is the thermal expansivity of water (assumed to be constant at $6.24 \times 10^{-4} \text{ K}^{-1}$, a typical value at 100 °C, 30 MPa), α_m is the mineral thermal expansivity (assumed to be constant at

¹Géosciences Azur CNRS, Université de Nice - Sophia Antipolis, 250 rue A. Einstein, 06560 Valbonne, France. ²Division of Earth and Planetary Sciences, Graduate School of Science, Kyoto University, Kyoto 606-8502, Japan.

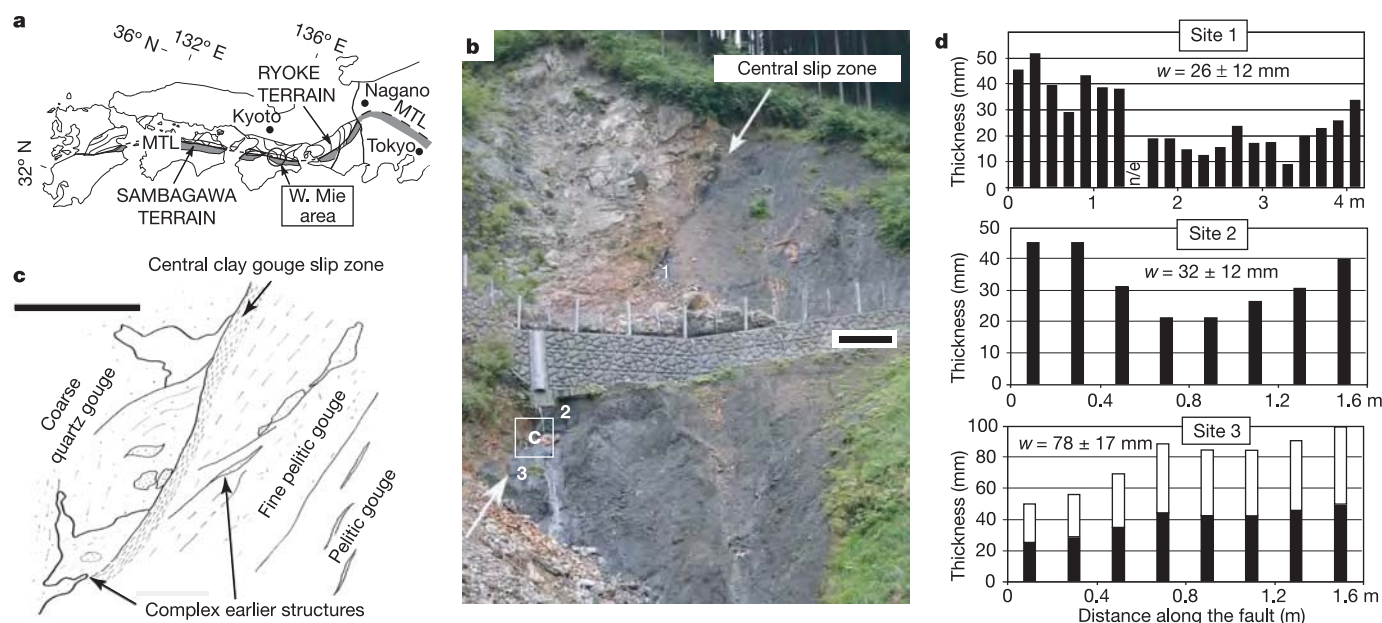


Figure 1 | The nature of a localized slip zone in a large geological fault. **a**, The Median Tectonic Line (MTL) in southwestern Japan. **b**, View facing east of the MTL fault zone in western Mie (scale bar, 2 m). **c**, Sketch of the details of the slip zone and its vicinity (scale bar, 0.3 m). **d**, Variation in slip

zone widths w along the exposed slip zone in three different sites (shown in **b**), with average and one standard deviation shown for each site. At site 3, the gradational nature of one edge is indicated by the white part of the bar.

$2 \times 10^{-5} \text{ K}^{-1}$) and α_{sf} is the porous medium thermal expansivity (assumed to be constant at $1 \times 10^{-5} \text{ K}^{-1}$) (ref. 9).

The laboratory-constrained parameters above were used to model the release of accumulating fluid pressure due to fluid flow from the slip zone and/or pore dilatancy during frictional heating. The fluid flow modelling is one-dimensional in the direction normal to the fault, and assumes behaviour to be constant along the fault. Calculations of thermal diffusion in slip zones suggest that an assumption of no heat loss from the slip zone is valid for timescales of the duration of an earthquake⁸. Negligible loss of fluid from the slip zone can also be assumed if the hydraulic diffusion length scale, L_h , is less than half of the width (w) of the slip zone, with $L_h = D_h \Psi$, where Ψ , a characteristic time controlling the rate of thermal pressurization^{8,9}, is:

$$\Psi = \frac{\rho c w}{\mu_d \Gamma_{P_e, P_f, T} V} \quad (4)$$

Equation (4) assumes isovolumetric shear at constant shear strain rate, where ρ is density (assumed constant at $2,600 \text{ kg m}^{-3}$), c = specific heat capacity ($1,000 \text{ J kg}^{-1} \text{ K}^{-1}$), μ_d is the dynamic coefficient of friction (measured at 0.4 for the slip zone gouge) and V is relative slip velocity.

Figure 3 presents the hydraulic diffusion lengths calculated using conditions of T , P_{lith} and P_f expected at depths of 2 and 6 km, based on the laboratory-constrained parameters. The effect of the rise in permeability and storage capacity with increasing fluid pressure (decreasing P_e) during the thermal pressurization process is illustrated by the different $\lambda = P_f/P_{\text{lith}}$ ratios presented in Fig. 3. Figure 3 shows that at a slip rate of 1 m s^{-1} , negligible fluid loss can be assumed (from the condition $L_h < 0.5w$) provided that w is not less than around $100 \mu\text{m}$ at 6 km depth and 1–2 mm at 2 km depth, which was never observed at the outcrop of the exhumed MTL fault zone (Fig. 1). The condition also holds at slip rates as low as 0.1 m s^{-1} , except perhaps for depths as shallow as 2 km where slip zone widths greater than 10–20 mm would be required for the assumption of negligible fluid loss to be valid (Fig. 3). These results indicate that the clay-rich gouge slip zone studied has a sufficiently low hydraulic diffusivity to trap pressurized fluid in the slip zone during rapid slip.

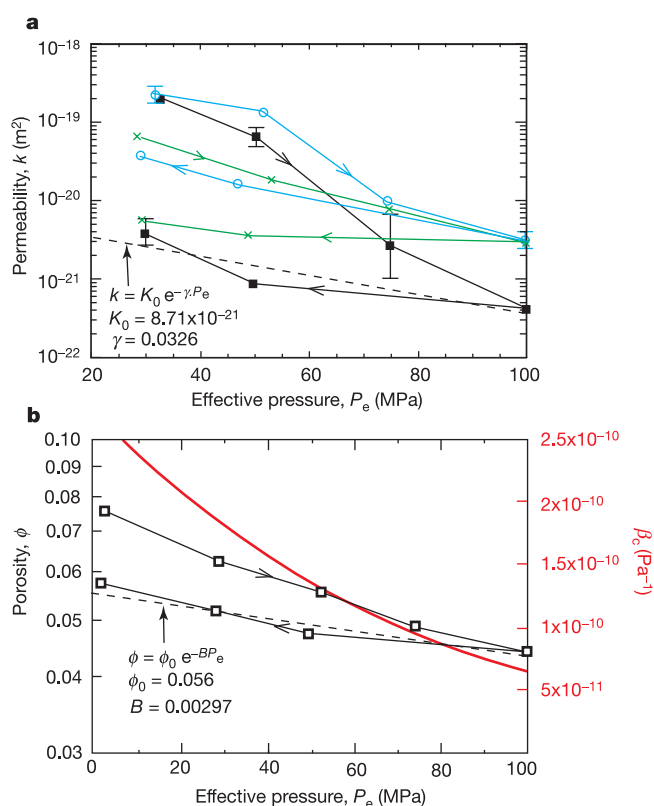


Figure 2 | Permeability and porosity data for slip zone rocks. Measurement error bars (one standard deviation) shown when larger than the symbol. **a**, Permeability data (pore pressure oscillation technique), for cores sampled perpendicular to the fault (black), parallel to fault and slip lineations (green), and parallel to fault and perpendicular to slip lineations (blue). Dashed line: best-fit log-linear equation for permeability perpendicular to the fault during decreasing effective pressure. **b**, Porosity data (black) and the resultant calculated storage capacity curve, β_c (red) from the best-fit log-linear porosity equation (black dashed line) and compressibility data cited in the text.

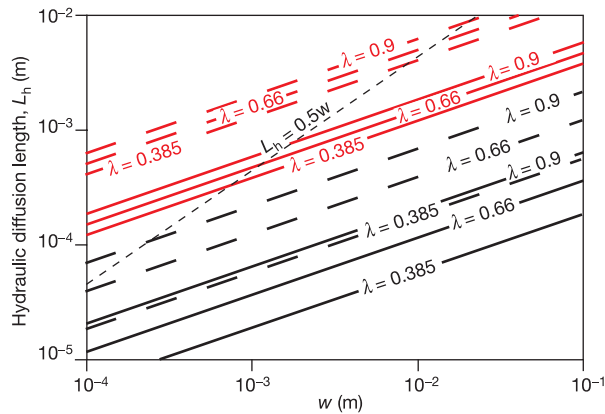


Figure 3 | Hydraulic diffusion lengths during thermal pressurization as a function of slip zone width. Cases are shown for depths of 2 km (red) and 6 km (black) and for slip rates of 1 m s^{-1} (solid lines) and 0.1 m s^{-1} (dashed lines). λ values (ratios of fluid pressure to lithostatic pressure) represent different stages of fluid pressurization, starting at hydrostatic pressure ($\lambda = 0.385$). Note that L_h is always smaller than $0.5w$ except for small w , shallow depths and low slip rates.

In the case of no fluid loss suggested above, the shear stress τ required to continue driving high-velocity fault slip at constant slip rate (we assume $V = 1 \text{ m s}^{-1}$) is related to normal stress (σ_n), initial fluid pressure (P_{f0}) and post-acceleration slip displacement d (refs 8, 9):

$$\tau_d = \mu_d(\sigma_n - P_{f0}) \exp\left(-\frac{d}{D_c}\right) \quad (5)$$

where $D_c = w\Psi_e$ is the characteristic slip distance of thermal pressurization (that is, the slip distance over which resistance drops to e^{-1} of its initial value), and

$$\Psi_e = \frac{\rho c}{\mu_d \Gamma} \quad (6)$$

controls the amount of weakening per unit shear strain in the slip zone. This analysis assumes instantaneous acceleration, a reasonable approximation for propagation sites and continued slip over the majority of the rupture surface, but possibly less appropriate to the initial nucleation zone where acceleration is slower. Figure 4 shows that D_c varies from about 0.03 to 0.5 m, in the ranges of slip zone width and depth studied. A narrower slip zone and a greater depth result in a faster rate of frictional heating and hence a smaller D_c (Fig. 4). These predicted slip-weakening distances are close to those inferred from seismological data for large earthquakes^{4–6} if the width of the deformation zone is smaller than about 50 mm as recognized here for several faults. The thermal pressurization process is therefore a high-velocity weakening mechanism that can solve the slip-weakening paradox quantitatively²⁵.

These results have important implications for the origin of seismic asperities^{26,27}. Earthquake instabilities arise from a force imbalance if the fault weakens at a faster rate than the drop in driving shear stress in the surrounding medium during rupture. Hence, for similar amounts of strength drop, fault motion is expected to be less stable for smaller D_c . The lateral continuity of the narrow central slip zone, and its low permeability, are evident on a scale of tens of metres, as shown in this paper. However, work describing the MTL fault zone approximately 200 km along strike did not show such a low permeability in the central portions of the fault, but a wider (2 m) zone of moderate-permeability gouge²⁸. The large-scale spatial heterogeneity in the slip zone width and its permeability can result in large variations in the degree and rate of slip weakening by thermal pressurization and thereby generate a high degree of heterogeneity in stress drop on the fault surface. If thermal pressurization is the

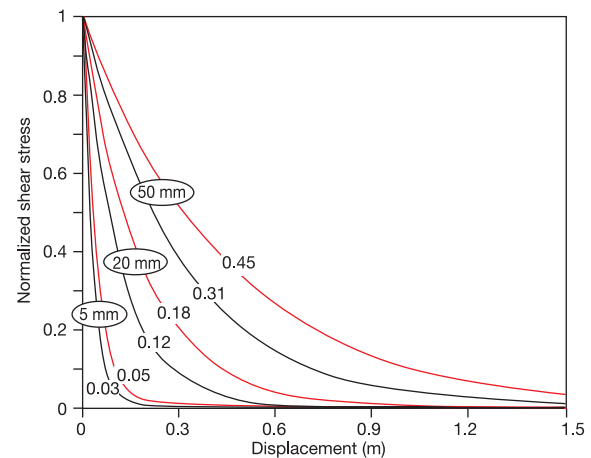


Figure 4 | Shear stress modelling for the thermal pressurization process. Shear stress evolution with slip in the weakening regime is normalized to the starting value, modelled at $V = 1 \text{ m s}^{-1}$ for depths of 2 km (red) and 6 km (black). We assumed an initial hydrostatic fluid pressure and lithostatic normal stress for the MTL, although this fault is a strike-slip fault and the normal stress acting on it cannot be estimated uniquely. Results are shown for three different slip zone widths as indicated in ovals. Numbers on the curves indicate the calculated D_c values (in metres).

predominant dynamic weakening mechanism, a large and straight fault segment with a thin slip zone of low permeability is most likely to be the site of a dynamic seismic asperity. Less effective thermal pressurization in a zone of diffuse deformation and/or high permeability could provide a mechanism for stress-roughening slip events, or 'barrier'-type heterogeneities. Asperities and barriers have been considered mostly as strong patches along faults^{26,27}, yet high strength does not necessarily warrant unstable fault motion. So although there are other possible high-velocity slip-weakening mechanisms^{29,30}, thermal pressurization may turn out to be the most significant mechanism besides geometric fault barriers in inducing heterogeneity in dynamic fault motion.

Received 3 March; accepted 7 June 2005.

- Jaeger, J. C. & Cook, N. G. W. *Fundamentals of Rock Mechanics* (Chapman & Hall, London, 1979).
- Dieterich, J. H. Time-dependent friction and the mechanics of stick slip. *Pure Appl. Geophys.* **116**, 790–806 (1978).
- Ohnaka, M. Earthquake source nucleation: a physical model for short-term precursors. *Tectonophysics* **211**, 149–178 (1992).
- Ide, S. & Takeo, M. Determination of constitutive relations of fault slip based on seismic wave analysis. *J. Geophys. Res.* **102**, 27379–27391 (1997).
- Mikumo, T., Olsen, K. B., Fukuyama, E. & Yagi, Y. Stress-breakdown time and slip-weakening distance inferred from slip-velocity functions on earthquake faults. *Bull. Seismol. Soc. Am.* **93**, 264–282 (2003).
- Fukuyama, E., Mikumo, T. & Olsen, B. Estimation of the critical slip-weakening distance: theoretical background. *Bull. Seismol. Soc. Am.* **93**, 1835–1840 (2003).
- Sibson, R. H. Interactions between temperature and pore-fluid pressure during earthquake faulting and a mechanism for partial or total stress relief. *Nature* **243**, 66–68 (1973).
- Lachenbruch, A. Frictional heating, fluid pressure, and the resistance to fault motion. *J. Geophys. Res.* **85**, 6097–6112 (1980).
- Mase, C. W. & Smith, L. Effects of frictional heating on the thermal, hydrological, and mechanical response of a fault. *J. Geophys. Res.* **92**, 6249–6272 (1987).
- Andrews, D. J. A fault constitutive relation accounting for thermal pressurization of pore fluid. *J. Geophys. Res.* **107**, doi:10.1029/2002JB001942 (2003).
- Ida, Y. Cohesive force across the tip of a longitudinal shear crack and Griffith's specific surface energy. *J. Geophys. Res.* **77**, 3796–3805 (1972).
- Ohnaka, M. & Shen, L.-F. Scaling of the rupture process from nucleation to dynamic propagation: implications of geometric irregularity of the rupturing surfaces. *J. Geophys. Res.* **104**, 817–844 (1999).
- Dieterich, J. H. in *Mechanical Behaviour of Crustal Rocks* (eds Carter, N. L., Friedman, M., Logan, J. M. & Stearns, D. W.) 103–120 (Geophysical

- Monograph 24, American Geophysical Union, Washington DC, 1981).
14. Scholz, C. H. The critical slip distance for seismic faulting. *Nature* **336**, 761–763 (1988).
 15. Marone, C. J. & Kilgore, B. D. Scaling of the critical slip distance for seismic faulting with shear strain in fault zones. *Nature* **362**, 618–621 (1993).
 16. Sibson, R. H. Generation of pseudotachylite by ancient seismic faulting. *Geophys. J. R. Astron. Soc.* **43**, 775–794 (1975).
 17. Fukuchi, T. Strong ferromagnetic resonance signal and magnetic susceptibility of the Nojima pseudotachylite in Japan and their implication for coseismic electromagnetic changes. *J. Geophys. Res.* **108**, 2312, doi:10.1029/2002JB002007 (2003).
 18. Wibberley, C. A. J. & Shimamoto, T. Internal structure and permeability of major strike-slip fault zones: the Median Tectonic Line in W. Mie Prefecture, S.W. Japan. *J. Struct. Geol.* **25**, 59–78 (2003).
 19. Sibson, R. H. Thickness of the seismic slip zone. *Bull. Seismol. Soc. Am.* **93**, 1169–1178 (2003).
 20. Suwa, K., Miyakawa, K., Mizutani, S., Hayashida, M. & Ohiwa, Y. Tsukide outcrop: a large outcrop of the Median Tectonic Line, central Kii Peninsular, Japan. *J. Geol. Soc. Jpn* **103**, 35–36 (1997).
 21. Tsutsumi, A. *et al.* Principal fault zone width and permeability of the active Neodani fault, Nobi active fault system, Southwest Japan. *Tectonophysics* **379**, 93–108 (2004).
 22. Blanpied, M. L. & Beeler, N. M. in *Fault-Related Rocks: A Photographic Atlas* (eds Snoke, A., Tullis, J. & Todd, V. R.) 72–73 (Princeton Univ. Press, Princeton, 1998).
 23. Faulkner, D. R. & Rutter, E. H. The gas-permeability of clay-bearing fault gouge at 20 °C. In *Faulting, fault sealing and fluid-flow in hydrocarbon reservoirs* (eds Knipe, R. J., Jones, G. & Fischer, Q.) 147–156, (1998).
 24. Wibberley, C. A. J. Hydraulic diffusivity of fault gouge zones and implications for thermal pressurization during seismic slip. *Earth Planets Space* **54**, 1153–1171 (2002).
 25. Wibberley, C. A. J. Hydrodynamical weakening of granular fault gouge during earthquake slip. *Geophys. Res. Abstr.* **6**, 02051 (2004).
 26. Kanamori, H. in *Earthquake Prediction: an International Review* (eds Simpson, D. W. & Richards, P. G.) 1–19 (Maurice Ewing Series 4, American Geophysical Union, Washington DC, 1981).
 27. Aki, K. Asperities, barriers, characteristic earthquakes and strong motion prediction. *J. Geophys. Res.* **89**, 5867–5872 (1984).
 28. Uehara, S. & Shimamoto, T. Gas permeability evolution of cataclasite and fault gouge in triaxial compression and implications for changes in fault-zone permeability structure through the earthquake cycle. *Tectonophysics* **378**, 183–195 (2004).
 29. Hirose, T. & Shimamoto, T. Fractal dimension of molten surfaces as a possible parameter to infer the slip-weakening distance of faults from natural pseudotachylites. *J. Struct. Geol.* **25**, 1569–1574 (2003).
 30. Di Toro, G., Goldsby, D. L. & Tullis, T. E. Friction falls towards zero in quartz rock as slip velocity approaches seismic rates. *Nature* **427**, 436–439 (2004).

Acknowledgements C.A.J.W. acknowledges financial support from the Japan Society for the Promotion of Science (JSPS), and assistance in the field from S. Uehara. This work was partially supported by a Grant-in-Aid for Scientific Research, JSPS, and by the Kyoto University 'Active Geosphere Project for the 21st Century'.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.A.J.W. (wibbs@geoazur.unice.fr).

Dental microwear texture analysis shows within-species diet variability in fossil hominins

Robert S. Scott^{1*}, Peter S. Ungar^{1*}, Torbjorn S. Bergstrom², Christopher A. Brown², Frederick E. Grine³, Mark F. Teaford⁴ & Alan Walker⁵

Reconstructing the diets of extinct hominins is essential to understanding the paleobiology and evolutionary history of our lineage. Dental microwear, the study of microscopic tooth-wear resulting from use^{1–4}, provides direct evidence of what an individual ate in the past. Unfortunately, established methods^{5–10} of studying microwear are plagued with low repeatability and high observer error¹¹. Here we apply an objective, repeatable approach for studying three-dimensional microwear surface texture to extinct South African hominins. Scanning confocal microscopy^{12,13} together with scale-sensitive fractal analysis^{14–19} are used to characterize the complexity and anisotropy of microwear. Results for living primates show that this approach can distinguish among diets characterized by different fracture properties. When applied to hominins²⁰, microwear texture analysis indicates that *Australopithecus africanus* microwear is more anisotropic, but also more variable in anisotropy than *Paranthropus robustus*. This latter species has more complex microwear textures, but is also more variable in complexity than *A. africanus*. This suggests that *A. africanus* ate more tough foods and *P. robustus* consumed more hard and brittle items, but that both had variable and overlapping diets.

To understand the relationships between diet and dental microwear, we first examined living animals with different diets. *Cebus apella*, for example, eats more fruit flesh and hard, brittle seeds², whereas *Alouatta palliata* consumes more leaves and other tough items^{22,23}. We measured microwear complexity by area-scale fractal complexity (Asfc) (Fig. 1). The Asfc is significantly higher ($\chi^2 = 36.97$, $P < 0.0001$; Kruskal-Wallis test) and more variable ($F = 81.65$, $P < 0.0001$; F -test) for *C. apella* (13.99 ± 11.034 , all values are mean $\pm$ s.d. unless otherwise stated) than for *A. palliata* (0.98 ± 1.221). The anisotropy variable, exact proportion length-scale anisotropy of relief (epLsar) (calculated with a scale observation of $1.8 \mu\text{m}$, epLsar_{1.8}), quantifies the degree of directionality in surface roughness at a fine scale. The anisotropy is significantly higher ($\chi^2 = 16.32$, $P < 0.0001$; Kruskal-Wallis test) and more variable ($F = 2.38$, $P < 0.05$) for *A. palliata* (epLsar_{1.8} 0.0052 ± 0.00215) than for *C. apella* (0.003 ± 0.0014) (Fig. 2a, b). These results suggest that hard, brittle foods associated with pits in microwear feature analyses^{9,10,24} leave a more complex microwear texture ($> \text{Asfc}$), whereas tough foods associated with scratches in microwear feature analyses^{9,10,24} produce a more anisotropic microwear texture ($> \text{Lsar}$).

Fossil hominin results indicate that *P. robustus* (Asfc 4.29 ± 2.150) has microwear textures more complex ($\chi^2 = 8.17$, $P < 0.005$; Kruskal-Wallis test) and more variable in complexity ($F = 16.82$,

$P < 0.0005$) than *A. africanus* (Asfc 1.686 ± 0.52) (Fig. 2c, d). These results are consistent with the hypothesis that *P. robustus* incorporated more hard and brittle foods in its diet^{20,25}. However, some overlap in Asfc for the hominins (Fig. 3b) suggests that *P. robustus* was unlikely to have been a specialized hard-object feeder. It is more

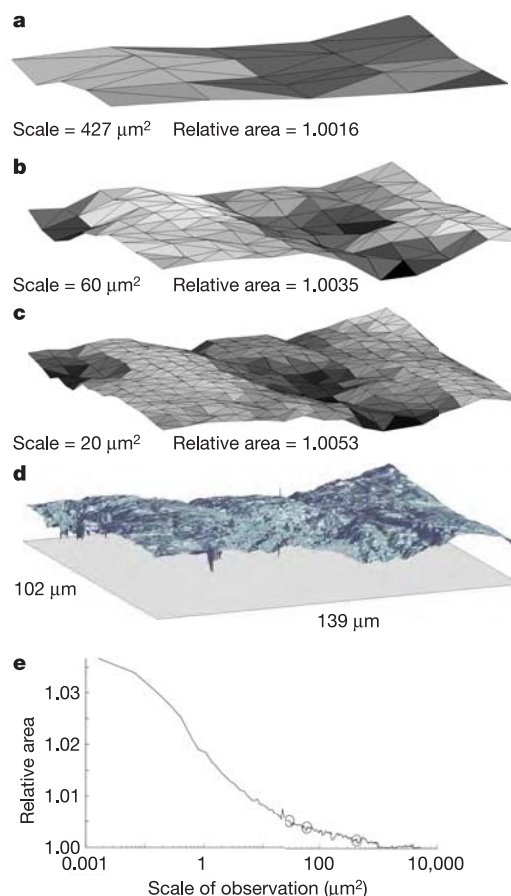


Figure 1 | Scale-sensitive fractal analysis. a–e, Relative area is calculated by dividing the area of a surface, calculated using triangles of a given scale in a virtual tiling algorithm (a, b, c), by the projected area of the surface (d). Relative area can then be plotted against scale in a log–log plot (e). Asfc³⁰ is a scale-sensitive measure of roughness and is the slope of the steepest part of the curve fitted to the plot of relative area over scale, multiplied by $-1,000$.

¹Department of Anthropology, University of Arkansas, Fayetteville, Arkansas 72701, USA. ²Surface Metrology Laboratory, Department of Mechanical Engineering, Worcester Polytechnic Institute, Worcester, Massachusetts 01609, USA. ³Departments of Anthropology and Anatomical Sciences, Stony Brook University, Stony Brook, New York 11794, USA. ⁴Center for Functional Anatomy & Evolution, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA. ⁵Departments of Anthropology and Biology, Pennsylvania State University, University Park, Pennsylvania 16802, USA.

*These authors contributed equally to this work.

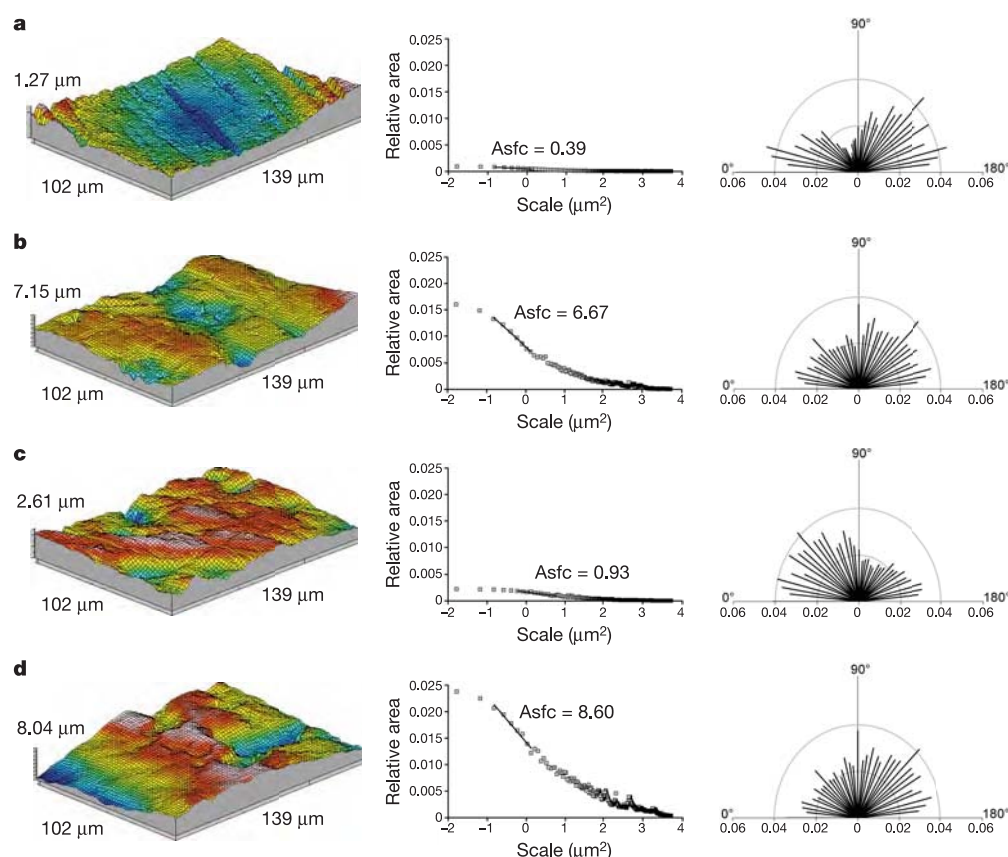


Figure 2 | Microwear texture analyses. **a–d**, Meshed axonometrics of digital elevation models (left), bivariate plots of relative area versus scale (middle), and rosette plots of normalized relative length at 1.8 μm by orientation (right) for scans on representative specimens of *Alouatta palliata* (NMNH 543117) (**a**), *Cebus apella* (NMNH 518433) (**b**), *Australopithecus africanus* (Sts 61) (**c**), and *Paranthropus robustus* (SK 16) (**d**). Steeper best-fit lines for the steepest order of magnitude on the relative area–scale curves evince higher Asfc values, indicating greater complexity (for example, **b** and **d**). More clumped rosettes (for example, **a** and **c**) have higher values of epLsar, indicating greater anisotropy of the wear fabric.

likely that hard, brittle foods were an occasional but important part of the diet. Previous studies have emphasized average differences between species rather than overlap, because low repeatability¹¹ associated with observer error made assessments of within-species variability difficult.

In contrast, the microwear textures of *Australopithecus africanus* (epLsar_{1.8} 0.0045 $\pm$ 0.00163) show greater anisotropy ($\chi^2 = 3.84$, $P = 0.05$; Kruskal–Wallis test) and epLsar variability ($F = 7.38$, $P < 0.01$) than *P. robustus* (epLsar_{1.8} 0.0028 $\pm$ 0.00060) (Fig. 2c, d). These data suggest a tougher diet on average for *A. africanus* compared with *P. robustus*, but one that is also more variable in its toughness.

Microwear texture analysis is a repeatable alternative to methods that rely on subjective feature counts and measurements on two-dimensional photomicrographs generated using a scanning electron microscope (SEM). Variables including Asfc and Lsar, which are based on scale-sensitive fractal analysis, can distinguish microwear found on molars of extant primates with contrasting diets (Fig. 3a). These variables also distinguish the microwear found on molars of extinct South African hominins (Fig. 3b). The reduction in observer error and the possibility of using larger samples dramatically improves estimations of microwear variability, and creates a new opportunity to model diet and subsistence variability in both extinct species and bioarchaeological samples.

The evidence presented here with respect to South African hominin molar microwear affirms some differences between *P. robustus* and *A. africanus*. For instance, wear fabrics of *P. robustus* are more complex and less anisotropic than those of *A. africanus*. These differences suggest a diet composed of more hard and brittle foods for *P. robustus* and more tough foods for *A. africanus*^{20,25}.

On the other hand, the contrasts in microwear texture variability found here offer new insights. The greater variation in complexity for *P. robustus* and in anisotropy for *A. africanus* suggests that these species altered different components of their diet, but that there was

probably substantial overlap in the fracture properties of their preferred foods. Thus, the clear differences between *A. africanus* and *P. robustus* microwear may relate, in part, to differences in critical dietary resources consumed only periodically during the year.

The study of dietary variability in fossil hominins has until now been problematic, given small sample sizes, subjective identification of individual features and observer error¹¹. The overlap and variability identified here for *Paranthropus* and *Australopithecus* was not apparent from earlier studies^{20,25}. This suggests that early hominin diet differences might relate more to microhabitat, seasonality or fall-back food choice than to oversimplified, dichotomous food preferences^{26–28}.

METHODS

The samples of the extant primates included high-resolution epoxy replicas of mandibular second molars of the New World monkeys, *Cebus apella* ($n = 35$) and *Alouatta palliata* ($n = 25$). Specimens are accessioned at the US National Museum of Natural History and provenience is known only generally. *C. apella* specimens were collected in central and eastern Brazil and *A. palliata* specimens were collected in Panama (from Darien to Bocas del Toro). The fossil specimens were the same as those considered by Grine²⁵, including *Paranthropus robustus* from Swartkrans (~ 1.9 – 1.5 Myr ago; $n = 9$) and *Australopithecus africanus* from Sterkfontein (~ 2.8 – 2.4 Myr ago; $n = 10$).

Specimens were examined using a Sensofar Plμ white-light scanning confocal imaging profiler with a $\times 100$ objective^{12,13}. Four adjacent areas on Facet 9 (ref. 29) of the specimens were scanned, sampling a total area of a $276 \times 204 \mu\text{m}$. These were then levelled using SolarMap Universal, producing digital elevation models with a vertical sampling interval of $0.005 \mu\text{m}$ and a lateral (x and y) sampling interval of $0.18 \mu\text{m}$.

The resulting data were analysed using scale-sensitive fractal analysis^{14–19}, an objective and repeatable approach to quantifying the complexity and directionality of surface roughness and their scales. Scale-sensitive fractal analysis is based on the idea that the apparent area of a rough surface, and the length of a profile from a rough surface, change with the scale of observation. Thus, surface textures appear smooth at sufficiently coarse scales, and rough with increasing resolution

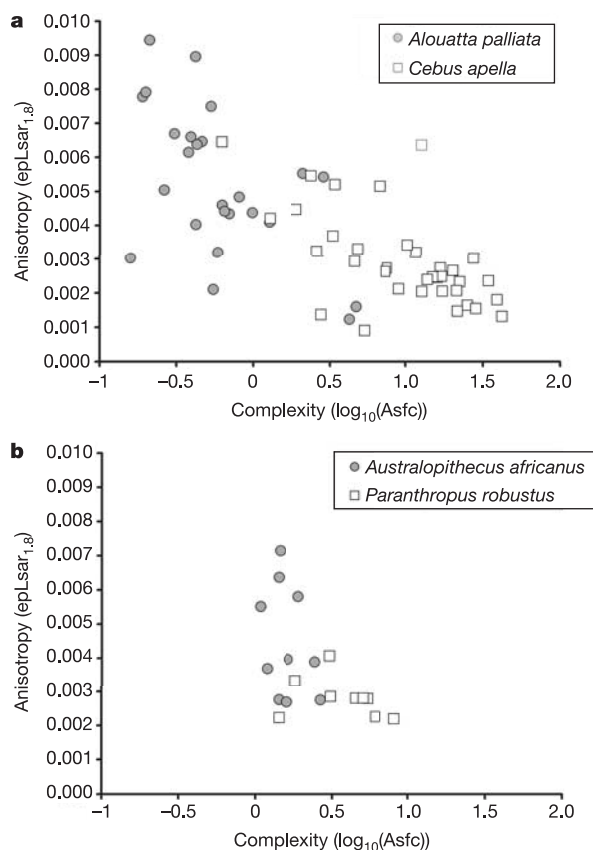


Figure 3 | Anisotropy and complexity. **a, b,** Bivariate plots of $epLsar_{1.8}$ versus $Asfc$ for *Alouatta palliata* and *Cebus apella* (**a**), and *Australopithecus africanus* and *Paranthropus robustus* (**b**). The values plotted are means based on the four scans with adjoining edges from each specimen.

at sufficiently fine scales. Changes in areas with scale are used to characterize complexity. These are calculated as changes in relative area, defined as the calculated surface area at a particular scale (Fig. 1a–c) divided by planometric area (Fig. 1d). The slope of the steepest part of the curve fitted to a log–log plot of relative area over some scale range (multiplied by $-1,000$) (Figs 1e and 2) is termed area–scale fractal complexity ($Asfc$)³⁰. For each scan studied here, 560 relative areas were calculated for scales from $\sim 5,300 \mu m^2$ to $\sim 0.17 \mu m^2$ using Kfrax (<http://www.surfract.com>). $Asfc$ was calculated for each scan over one order of magnitude, and mean $Asfc$ was calculated for each specimen.

Relative lengths of depth profiles differ with orientation when the roughness of a surface has directionality (that is, when the surface is anisotropic). We define relative lengths at given orientations as vectors. Thirty-six vectors were calculated at 5° intervals for each scale and then normalized. For a given scale, normalized vector lengths or ‘exact proportion’ $relL_a$ are equal to $(relL_a - 1) / \Sigma(relL_a - 1)$, where $relL$ is relative length and a is the orientation of the length scale analysis. This normalization is functionally equivalent to normalizing by the mean of $relL$ when the number of orientations calculated is constant across the comparison. Longer normalized relative length vectors in the rosette plots generally correspond to more wear features normal to the direction of the vector.

Normalized relative length vectors can be displayed graphically in a rosette diagram (Fig. 2). The length of the mean vector is a new measure of surface anisotropy called exact proportion length–scale anisotropy of relief ($epLsar$). $epLsar$ was calculated for each scan using Kfrax with a scale of observation of $1.8 \mu m$, and mean $epLsar$ was calculated for each specimen.

Descriptive and inferential statistics were computed using SAS 9.1.

Received 18 April; accepted 17 May 2005.

- Walker, A. C., Hoeck, H. N. & Perez, L. Microwear of mammalian teeth as an indicator of diet. *Science* **201**, 808–810 (1978).
- Rensberger, J. M. in *Development, Function and Evolution of Teeth* (eds Butler, P. M. & Joysey, K. A.) 415–438 (Academic, New York, 1978).
- Ryan, A. S. Anterior dental microwear and its relationship to diet and feeding behaviour in three African primates (*Pan troglodytes troglodytes*, *Gorilla gorilla*

gorilla, and *Papio hamadryas*). *Primates* **22**, 533–550 (1981).

- Gordon, K. D. A study of microwear on chimpanzee molars: implications for dental microwear analysis. *Am. J. Phys. Anthropol.* **59**, 195–215 (1982).
- Solounias, N., Teaford, M. & Walker, A. Interpreting the diet of extinct ruminants: the case of a non-browsing giraffid. *Paleobiology* **14**, 287–300 (1988).
- Ungar, P. S., Simons, J.-C. & Cooper, J. W. A semiautomated image analysis procedure for the quantification of dental microwear. *Scanning* **13**, 31–36 (1991).
- Pastor, R. F. Dietary adaptations and dental microwear in Mesolithic and Chalcolithic South Asia. *J. Hum. Ecol.* (special issue) **2**, 215–228 (1992).
- Molleson, T., Jones, K. & Jones, S. Dietary change and the effects of food preparation on microwear patterns in the Late Neolithic of abu Hureyra, northern Syria. *J. Hum. Evol.* **24**, 455–468 (1993).
- Teaford, M. F. A review of dental microwear and diet in modern mammals. *Scanning Microsc.* **2**, 1149–1166 (1988).
- Strait, S. G. Molar microwear in extant small-bodied faunivorous mammals: an analysis of feature density and pit frequency. *Am. J. Phys. Anthropol.* **92**, 63–79 (1993).
- Grine, F. E., Ungar, P. S. & Teaford, M. F. Error rates in dental microwear quantification using SEM. *Scanning* **24**, 144–153 (2002).
- Boyde, A. & Fortelius, M. New confocal LM method for studying local relative microrelief with special references to wear studies. *Scanning* **13**, 429–430 (1991).
- Ungar, P. S., Brown, C. A., Bergstrom, T. S. & Walker, A. Quantification of dental microwear by tandem scanning confocal microscopy and scale-sensitive fractal analyses. *Scanning* **25**, 189–193 (2003).
- Brown, C. A. & Savary, G. Describing ground surface texture using contact profilometry and fractal analysis. *Wear* **141**, 211–226 (1991).
- Brown, C. A., Charles, P. D., Johnsen, W. A. & Chesters, S. Fractal analysis of topographic data by the patchwork method. *Wear* **161**, 61–67 (1993).
- Brown, C. A., Johnsen, W. A. & Hult, K. M. Scale-sensitivity, fractal analysis and simulations. *Int. J. Mach. Tool. Manu.* **38**, 633–637 (1998).
- DeChiffre, L. et al. Quantitative characterization of surface texture. *Annals CIRP* **49**, 635–652 (2000).
- Brown, C. A. & Siegmund, S. Fundamental scales of adhesion and area-scale fractal analysis. *Int. J. Mach. Tool. Manu.* **41**, 1927–1933 (2001).
- Bergstrom, T. S. & Brown, C. A. Interaction between horizontal scanning instruments and surfaces. *Int. J. Mach. Tool. Manu.* **41**, 1995–2000 (2001).
- Grine, F. E. & Kay, R. F. Early hominid diets from quantitative image analysis of dental microwear. *Nature* **333**, 765–768 (1988).
- Eisenberg, J. F. *Mammals of the Neotropics* Vol. 1. *The Northern Neotropics: Panama, Colombia, Venezuela, Guyana, Suriname, French Guiana* 233–261 (Univ. Chicago Press, Chicago, 1989).
- Estrada, A. Resource use by howler monkeys (*Alouatta palliata*). *Int. J. Primatol.* **5**, 105–131 (1984).
- Neville, M. K., Glander, K. E., Braza, F. & Rylands, A. B. in *Ecology and Behavior of Neotropical Primates* Vol. 2 (eds Mittermeier, R. A., Rylands, A. B., Coimbra-Filho, A. F. & de Fonseca, G. A. B.) 349–453 (World Wildlife Fund, Washington DC, 1988).
- Crompton, R. H., Savage, R. & Spears, I. H. The mechanics of food reduction in *Tarsius bancanus*: hard-object feeder, soft-object feeder or both? *Folia Primatol.* **69**(suppl. 1), 41–59 (1998).
- Grine, F. E. Dental evidence for dietary differences in *Australopithecus* and *Paranthropus*. *J. Hum. Evol.* **15**, 783–822 (1986).
- Wrangham, R. W., Conklin-Brittain, N. L. & Hunt, K. D. Dietary response of chimpanzees and Cercopithecines to seasonal variation in fruit abundance. I. Antifeedants. *Int. J. Primatol.* **19**, 971–998 (1998).
- Stanford, C. B. & Nkurunungi, J. B. Behavioral ecology of sympatric chimpanzees and gorillas in Bwindi Impenetrable National Park, Uganda: Diet. *Int. J. Primatol.* **24**, 901–918 (2003).
- Lambert, J. E., Chapman, C. A., Wrangham, R. W. & Conklin-Brittain, N. L. The hardness of cercopithecine foods: implications for the critical function of enamel thickness in exploiting fallback foods. *Am. J. Phys. Anthropol.* **125**, 363–368 (2004).
- Kay, R. F. The evolution of molar occlusion in the Cercopithecidae and early Catarrhines. *Am. J. Phys. Anthropol.* **46**, 327–352 (1977).
- American Society of Mechanical Engineers, *National Standard B46.1: Surface Texture, Surface Roughness, Waviness and Lay* (American Society of Mechanical Engineers, New York, 2002).

Acknowledgements We are grateful to the curators at the US National Museum of Natural History, the Transvaal Museum and the University of the Witwatersrand for permission to study specimens in their care, and thank A. Pérez-Pérez for his help preparing the hominin replicas. This project was funded by the US National Science Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare competing financial interests: details accompany the paper on www.nature.com/nature. Correspondence and requests for materials should be addressed to P.S.U. (pungar@uark.edu).

LETTERS

Refractory periods and climate forcing in cholera dynamics

Katia Koelle^{1*}, Xavier Rodó^{2*}, Mercedes Pascual¹, Md. Yunus³ & Golam Mostafa³

Outbreaks of many infectious diseases, including cholera, malaria and dengue, vary over characteristic periods longer than 1 year^{1,2}. Evidence that climate variability drives these interannual cycles has been highly controversial, chiefly because it is difficult to isolate the contribution of environmental forcing while taking into account nonlinear epidemiological dynamics generated by mechanisms such as host immunity²⁻⁴. Here we show that a critical interplay of environmental forcing, specifically climate variability, and temporary immunity explains the interannual disease cycles present in a four-decade cholera time series from Matlab, Bangladesh. We reconstruct the transmission rate, the key epidemiological parameter affected by extrinsic forcing, over time for the predominant strain (El Tor) with a nonlinear population model that permits a contributing effect of intrinsic immunity. Transmission shows clear interannual variability with a strong correspondence to climate patterns at long periods (over 7 years, for monsoon rains and Brahmaputra river discharge) and at shorter periods (under 7 years, for flood extent in Bangladesh, sea surface temperatures in the Bay of Bengal and the El Niño–Southern Oscillation). The importance of the interplay between extrinsic and intrinsic factors in determining disease dynamics is illustrated during refractory periods, when population susceptibility levels are low as the result of immunity and the size of cholera outbreaks only weakly reflects climate forcing.

Host immunity has a key role in the nonlinear dynamics of infectious diseases. Acquired immunity and the replenishment of susceptible individuals are capable of generating interannual disease cycles in nonlinear mathematical models of disease lacking any environmental forcing⁵. Seasonality can interact with this natural frequency to produce oscillations of longer period or more complex patterns, including chaos⁶. Thus, interannual disease cycles can either arise intrinsically, by means of epidemiological dynamics including seasonality, or be driven extrinsically, for instance by interannual climate variability at frequencies similar to those of the response in incidence levels. No evidence has yet been obtained for a role of climate variability in interannual disease cycles with an approach that allows for the alternative explanation that they are simply produced intrinsically⁴. The few exceptions have so far been limited to phenomenological treatments of disease dynamics^{7,8}. We use a nonlinear population model that takes into account immunity and disease transmission to show a strong correspondence between cholera transmission and climate variability.

The temporal cholera data consist of monthly symptomatic cases from the rural region of Matlab, Bangladesh, from 1966 to 2002, obtained from a surveillance programme by the International Center for Diarrhoeal Disease Research (Fig. 1). The study area is located 40 km southeast of the capital Dhaka, and lies in the delta region of

the Ganges and the Brahmaputra rivers. We focus the analysis on the temporal variability of the predominant strain, El Tor, whose dynamics exhibit clear seasonal and interannual variability, with peaks in spring and late autumn⁹.

To separate the roles of intrinsic feedbacks from extrinsic (environmental) forcing, we extended a recently proposed disease model¹⁰ to incorporate both immunity from previous El Tor infections and the likely possibility of cross-immunity from previous infections caused by the Classical strain (see Methods). The model contains two equations. The first one is a nonlinear transmission equation of the form

$$I_{t+1} = \beta_t I_t^\alpha \left(\frac{S_t}{N_t} \right)^\gamma \varepsilon_t \quad (1)$$

where I_t is the number of El Tor-infected individuals, S_t is the number of individuals susceptible to El Tor, N_t is the total population size, and ε_t is a multiplicative noise term, all at time t . The exponents α and γ are used to incorporate deviations from the random mixing assumption¹¹. Pathogen transmission rate β_t is a key parameter that we specifically let vary in time, to represent the effect of extrinsic

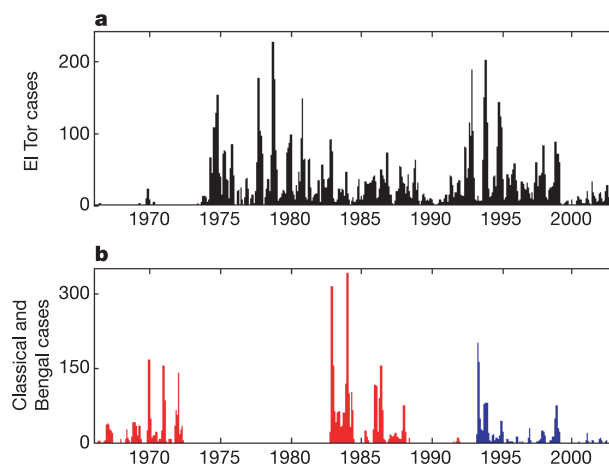


Figure 1 | Time series of cholera cases from 1966 to 2002, aggregated monthly. **a**, Cholera cases of the El Tor biotype. **b**, Cholera cases of the Classical biotype (red), and the Bengal strain (blue). Bengal cases were excluded before analysis because this strain belongs to a different serogroup (0139) from El Tor and Classical (01), and cross-immunity between serogroups seems to be absent²⁸. Population size was estimated using four census points and monthly demographic data adjusted for net emigration, and increased from 152,000 in 1966 to 223,000 in 2002 (not shown).

¹Department of Ecology and Evolutionary Biology, 2045 Kraus Natural Science Building, University of Michigan, 830 North University Avenue, Ann Arbor, Michigan 48109-1048, USA. ²ICREA and Climate Research Laboratory, Barcelona Science Park, University of Barcelona, c/Baldri i Reixach, 4-6 08028 Barcelona, Catalunya, Spain. ³International Center for Diarrhoeal Disease Research, Dhaka 1000, Bangladesh.

*These authors contributed equally to this work.

drivers on the transmission rate. This parameter corresponds to the number of contacts per infected individual per unit time, multiplied by the probability that a contact with an infected individual leads to infection¹². For cholera, contacts reflect faecal–oral transmission through food and water, and/or environmental transmission through contamination of aquatic environments used by humans. β_t contains both a seasonal component β_{seas} and a longer-term component β_{lt} , such that $\beta_t = \beta_{\text{seas}}\beta_{\text{lt}}$ (ref. 10). We divide β_t into two components to identify fluctuations in transmission rates that seasonality alone cannot explain. In particular, this formulation allows for mechanisms that influence interannual variability in β_t as the result of a modulation of the seasonality in transmission rates. A second equation specifies the number of susceptible individuals at time t :

$$S_t = N_t - \sum_{i=0}^m (\kappa_i I_{t-i}) - \sum_{i=0}^m (\kappa_i^{\text{cl}} I_{t-i}^{\text{cl}}) \quad (2)$$

where N_t is the current population size at time t , $\sum_{i=0}^m (\kappa_i I_{t-i})$ is the total number of individuals recovered from an El Tor infection and immune to El Tor reinfection, and $\sum_{i=0}^m (\kappa_i^{\text{cl}} I_{t-i}^{\text{cl}})$ is the total number of individuals recovered from infection by the Classical biotype and immune to El Tor reinfection. The functions κ_i and κ_i^{cl} describe the decay of immunity, with the subscript i indicating the number of

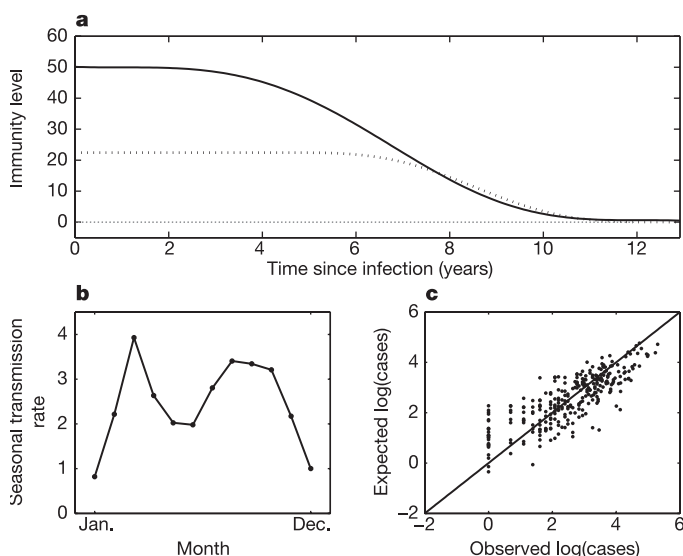


Figure 2 | Results of the nonlinear disease model. **a**, The decay of immunity functions κ_i and κ_i^{cl} for individuals having recovered from previous El Tor (solid line) and Classical (dotted line) infections, respectively. The decay of immunity quantifies the duration for which previously infected individuals are expected to be immune to reinfection with cholera of the El Tor biotype. The intercepts κ_i and κ_i^{cl} correspond to the ratio of asymptomatic to symptomatic infections¹⁰. This assumes that immunity is initially complete, a reasonable assumption given results from rechallenge studies¹⁴. The infection-to-case ratios estimated from these intercept values correspond to 50:1 for El Tor and 22:1 for Classical. These results fall within the published range of asymptomatic to symptomatic infection ratios for cholera²⁹ and are in agreement with the known higher asymptomatic ratio for El Tor than Classical³⁰. **b**, The seasonal component of the transmission rate, β_{seas} , anchored at its normalized December value β_{dec} , so that $\beta_{\text{dec}} = 1$. **c**, Logarithm of expected El Tor cases plotted against the logarithm of observed El Tor cases ($r^2 = 0.63$). The mixing exponent α equals 0.57 with γ set at 1 (Supplementary Information). The low value of α might reflect cholera's high degree of spatial clustering, a pattern noted empirically⁹. Alternatively, because *V. cholerae* temporarily survives in an aquatic reservoir, the dependence on the previous month's infections might be lower than for diseases with a strict direct transmission route. Dynamic simulations of the model without noise generate strict annual cycles. When simulated with dynamic noise, interannual variability does arise, but no regularities in dominant frequencies are found.

months since infection and the values of κ_i and κ_i^{cl} being the degrees of immunity that an individual has i months after being infected. Equations (1) and (2) are combined into a single expression relating incidence levels in the present to those in the past. The fit of this model to data relies on a semi-parametric approach^{10,13} because the long-term component of the transmission rate (β_{lt}), which incorporates any trends or fluctuations of periodicity longer than seasonal, is not specified or constrained in any way. The model itself reconstructs the patterns of time variation in transmission and associated susceptible levels. It further provides an estimate of the immunity and cross-immunity functions. Acquired immunity to *Vibrio cholerae* is known to exist, but its duration is highly debated¹⁴.

Results show that acquired immunity to reinfection with El Tor, from previous Classical and El Tor infections, is long-lasting (Fig. 2a). The degree of immunity from a previous El Tor infection starts to wane 3 years after infection, but partial immunity lasts for up to 10 years. Classical infections confer complete cross-immunity for more than 6 years, and full susceptibility to El Tor reinfection does not occur until more than 10 years after infection. A shorter duration of El Tor immunity than Classical cross-immunity agrees with results from field studies¹⁵, and may result from the greater severity of Classical infections producing stronger immunological responses. The intercepts of the immunity functions can be interpreted as providing estimates of ratios of asymptomatic to symptomatic cases (Fig. 2a). The higher ratio for El Tor than for Classical infection agrees with previous epidemiological studies (Fig. 2, legend). Seasonal transmission shows a clear monthly variation (Fig. 2b) consistent with the known seasonality of cases. More significantly, from the immunity curves, the fraction of the population susceptible to El Tor infection (the ratio of susceptible to total individuals; S/N) over time can be computed from equation (2) (Fig. 3b). Concurrent changes in the long-term transmission rate (Fig. 3a) provide evidence for the forcing of cholera by extrinsic factors at interannual timescales. The fit of the full model accounts for 63% of the variability in the logarithm of the El Tor cases (Fig. 2c). Additional effects of interannual forcing might still be contained in the residuals of the model, a point to which we return later when addressing specific environmental covariates.

We can now look more closely at the distinct roles of extrinsic (environmental) drivers and intrinsic (immunity) factors in the dynamics of cholera. Figure 3 shows the long-term transmission rate temporally aligned with the fraction of the host population susceptible to El Tor infection, as well as the cholera case data aggregated monthly. In times of high transmission, the response in El Tor cases would be expected to be large if intrinsic factors were

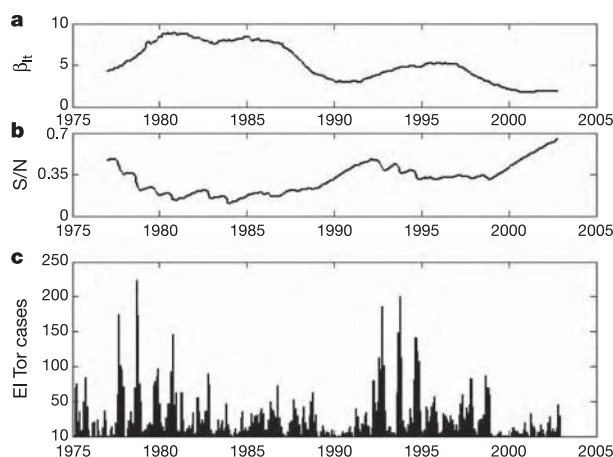


Figure 3 | Cholera refractory periods. **a**, The long-term component of the transmission rate, β_{lt} . **b**, The fraction of the population susceptible to cholera (S/N) over time. **c**, Time series of El Tor cases from 1976 to 2002.

unimportant. However, the magnitude of the response is only high at a subset of these times. Most notably, El Tor cholera cases are low for 1982–87, despite the favourable environmental conditions for high transmission of pathogens. This episode was concurrent with a low proportion of susceptible hosts after the higher incidence levels of the late 1970s and early 1980s (including the large Classical outbreak in 1982–83). Moreover, one of the highest El Tor cholera case episodes (1993–95) occurred when transmission rate was only moderately high but the susceptible fraction had had significant time to replenish itself (1988–92). These results illustrate that high transmission rates result in high case responses only when the immunity levels of the host population are low; that is, when the host population is not in a refractory period from previous disease outbreaks.

We now turn to the interpretation of changes in interannual patterns of transmission by examining associations with potential climatic drivers. These associations provide support for the model itself and for the assumption that the reconstructed variability in transmission reflects extrinsic forcing. We focus on rainfall, associated river discharge, and flood extent, because of the prominent role of water levels in the proposed mechanisms for cholera seasonality¹⁶. In the bimodal seasonal cycle of cholera cases (and also of transmission rates; Fig. 2b) there is a marked decrease during the summer monsoons, probably resulting from a reduction in cholera's environmental concentration and/or the decrease in salinity affecting its survival. Cholera cases increase again and peak with a lag after this season, as floods presumably concentrate the population on the decreased land area available and break down sanitary conditions, promoting secondary transmission through the more direct faecal–oral route. These two seasonal mechanisms of opposite effect at different lags can therefore mediate an influence of either positive or negative rainfall anomalies at interannual timescales. Besides rainfall, we also consider two remote drivers of interannual climate variability in the region, the El Niño–Southern Oscillation (ENSO) and sea surface temperatures (SSTs) in the Bay of Bengal, previously proposed to influence cholera in Bangladesh^{8,17}. Figure 4a shows a clear inverse relationship between the reconstructed transmission rate β_{lt} and both the low-frequency variation of rainfall from northeast India¹⁸ and Brahmaputra river discharge data, providing the first evidence for the long-standing hypothesis that rainfall and associated water levels drive cholera patterns. A long-term decreasing trend in transmission rate, associated with an opposite long-term trend in the Brahmaputra's discharge, is also evident. These environmental time series components were obtained by separating signal from noise and isolating the variability with a period longer than 7 years (see Methods). This period was chosen to focus initially on interannual variability at scales longer than those relevant for ENSO.

A second curve is shown in Fig. 4a for the long-term component of the transmission rate that closely matches the estimated β_{lt} . This curve was obtained by first aggregating the logarithm of the fitted long-term transmission rate, $\log \beta_{lt}$, with the residuals from the model, $\log \epsilon_t$. This non-seasonal transmission term allows us to consider all possible influences on incidence levels that were not accounted for by the epidemiological dynamics and seasonal forcing. The low-frequency component of transmission was then isolated from this aggregated variable, by using the same procedure as for the environmental variables. This approach produces a curve extremely similar to that originally fitted by the model. Interestingly, additional frequency components of this aggregated variable can now be isolated at shorter timescales, particularly those relevant to ENSO. Figure 4b shows the signal extracted for variability at periods less than 7 years superimposed on the SST anomalies averaged for the Niño3.4 region in the Pacific, and SST anomalies averaged for the Bay of Bengal. The SST in the Bay of Bengal lags the SST anomalies for the Niño3.4 region (Fig. 4b) by 2–3 months, establishing a connection between ENSO events and regional climate in Bangladesh, which we discuss below. In particular, there is a strong positive lagged correlation of this faster component of transmission with the 1987–88 and

the 1997–98 El Niño events. Locally, extreme floods in Bangladesh (Fig. 4b) and positive rainfall anomalies (Supplementary Fig. S3) occur during these ENSO years. A positive association between SST in the Bay of Bengal and this component of transmission is also evident. The 1987–88 ENSO episode is of special relevance because, despite the increase in transmission rate, the response in terms of cases occurs but is a weak one (Fig. 3c). This episode falls during the refractory period, for which the susceptible population is small and not yet ready to respond to increases in transmission rate. Only the

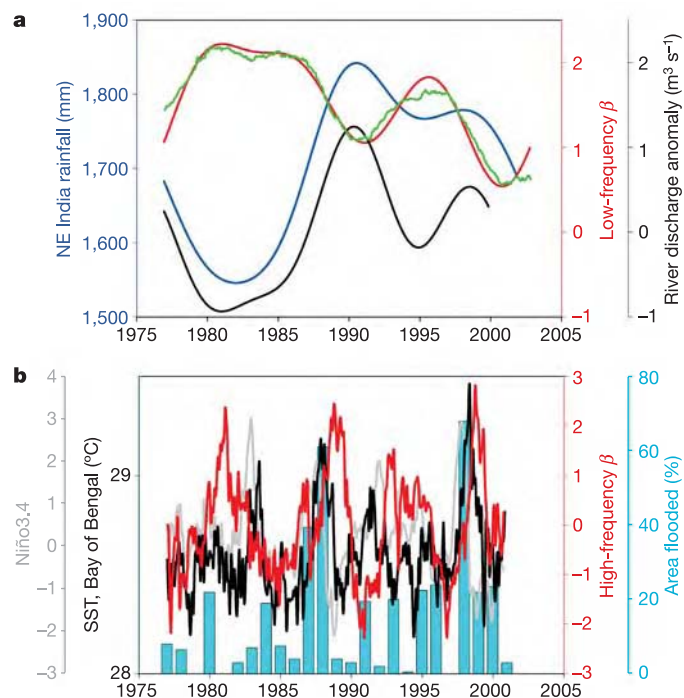


Figure 4 | Environmental drivers and transmission. The non-seasonal component of the transmission rate (green; plotted as $\log \beta_{lt}$) was added to the residuals of the model, and decomposed into frequencies of low and high variability (accounting for 27% and 57%, respectively, of the variance in this aggregated term). **a**, The low-frequency variation in the non-seasonal transmission rate (red), plotted with the low-frequency variation of northeast (NE) India rainfall (blue) and Brahmaputra river discharge (black). NE India rainfall covers an area of four subdivisions and $267 \times 10^3 \text{ km}^2$. As expected, the low-frequency variation of the NE India rainfall is reflected in the Brahmaputra river discharge variation ($r = +0.93$, $P < 0.05$, lag = 0 months). The low-frequency variation in transmission is negatively correlated with both the NE India rainfall ($r = -0.797$, $P < 0.05$, lag = 14 months) and Brahmaputra river discharge anomalies ($r = -0.9278$, $P < 0.02$, lag = 7 months). **b**, The high-frequency variation in the non-seasonal transmission rate (red), plotted with Niño3.4 (grey) and SST in the Bay of Bengal (black). SSTs in the Bay of Bengal are obtained from a 0° to 23° N and 80° to 100° E grid, with $2^\circ \times 2^\circ$ resolution from the extended reconstructed SSTs of the National Oceanic and Atmospheric Administration National Climate Data Center. Local correlations between Niño3.4 and SST in the Bay of Bengal reach +0.86, with a 2–3-month lag ($P < 0.05$). Local correlations between the high-frequency component of transmission and Niño3.4 reach maximal values between +0.63 and +0.91 at 8–10-month lags ($P < 0.05$) for 1986–87, 1990–91, 1994 and 1997–98. Similarly, local correlations between this transmission component and SST in the Bay of Bengal range between +0.68 and +0.96 for 0–9-month lags ($P < 0.05$) for 1982, 1986–87, 1988–89, 1990–91 and 1997. Bars show the percentage of the area of Bangladesh flooded over this time period. Annual flood area data come from the flood forecasting and warning centre (BWDB, Dhaka, Bangladesh). An area is considered flooded if it experiences at least one flooding event within the year's monsoon season. Local correlations between SST in the Bay of Bengal and NEIR (northeast India rainfall) (not shown) are negative, exceeding -0.85 at interannual timescales for periods longer than 7 years ($P < 0.0001$).

episode of 1982–83, for which the response is predominantly Classical and not El Tor, shows no increase in transmission rate. A time-series method that specifically quantifies the strength of the association between variables locally in time (see Methods) reveals that only one particular biotype at a time is responsible for the response of cases to climate for a given El Niño event (Supplementary Fig. S1).

The comparison of regional rainfall anomalies for Bangladesh between the different ENSO events reveals that 1982–83 exhibited drought conditions that were unique for this period in their intensity and spatial extent (Supplementary Figs S2 and S3). This drought provides an explanation for the major outbreak of Classical infection in place of El Tor. Anomalously low water levels (and associated higher salinity) have been proposed to favour this biotype relative to El Tor in spatial observations of biotype distributions in Bangladesh, with Classical found in more coastal regions¹⁹.

Thus, both floods and droughts seem to promote cholera transmission, depending on the strain and the temporal scale of interannual variability. This finding is consistent with the two different ways in which water levels have been postulated to influence the seasonality of cholera. This complex nonlinear response to water levels should be examined further with mechanistic mathematical models that couple cholera seasonality to its interannual variation. Our findings on SST in the Bay of Bengal (Fig. 4b) further support a role of rainfall and clarify the regional influence of climate variability on cholera at ENSO timescales. Two areas for SST in the Indian Ocean appear most relevant to the variability in Bangladesh rainfall, namely the central Indian Ocean²⁰ and the Bay of Bengal. SSTs in these regions exert a (nonlinear) influence on the subsequent summer monsoon²¹. These areas also appear strongly linked to El Niño²² (Fig. 4b for Bay of Bengal), although the degree of their dependency, and that of the so-called Indian Ocean dipole mode, on ENSO remain controversial²³. In addition to rainfall, water temperature in ponds and rivers provides another local mechanism for the remote association of cholera transmission with SST in the Bay of Bengal and the Pacific (ENSO). Changes in cloud cover, wind stress and evaporation modulate variations in the net heat flux entering the system^{22,24}, increasing both the SST in the Bay of Bengal and affecting the surface temperature over land. The resulting warming of water temperature in ponds and rivers might increase the incidence of cholera through the faster growth rate of the pathogen in aquatic environments.

We have shown the existence of refractory periods during which climate-driven increases in transmission do not result in large outbreaks. Once the interplay of climate forcing and disease dynamics is taken into account, clear evidence emerges for a role of climate variability in the transmission of cholera. A nonlinear population model achieves this by explicitly taking into account epidemiological dynamics: changes in the abundance of susceptible hosts, rates of decay of immunity and seasonal transmission could all be reconstructed from time series of infected individuals and total population. The finding of a high susceptible fraction in the Matlab population for recent times is of particular concern for the near future. Although a prolonged period of low transmission rate has been present since 2000, the system seems ripe for an outbreak if this rate were to increase. Future work should explicitly compare the predictability of models based solely on climate variables with that of models including intrinsic disease dynamics^{8,25}. Our results suggest that forecasting schemes will require the consideration of both climate variability and the fraction of susceptible individuals in the population. Given that time series on susceptible levels are rarely available, approaches such as that presented here are crucial in the attempt to forecast and anticipate the future size of outbreaks. Real-time monitoring on the state of the oceanic regions we have outlined will also be important for an effective early warning system based on climate, but will need to be integrated with susceptibility levels in future work.

METHODS

Fitting the extended population model. The extended model seeks to recover the seasonal transmission rates, the long-term transmission rates, the mixing exponent α and the decay of immunity functions κ and κ^{cl} , given only time series of cholera cases and population size. As in the original nonlinear population model¹⁰, the transmission equation is logarithmically transformed, making it a semi-parametric additive model. Using a backfitting algorithm, the parametric part of the model is first fitted through a combination of weighted least-squares regressions and recursive Taylor expansions. The backfitting algorithm is necessary to provide a progressive improvement in our estimate of the susceptible fraction in the population through the adjustment of the immunity functions. The nonparametric part of the model—that is, the long-term transmission rate—is then obtained by smoothing the residuals of the regression step. Two parameters (smoothing bandwidth h and spline penalty weight μ) determine the flexibility of the model and are objectively selected by cross-validation (see also Supplementary Information). Reported results shown in Figs 2 and 3 have $h = 23$ and $\mu = 10^{10}$. The extension of the model to two strains resides in the regression step, where the number of susceptible individuals is expressed as equation (2) for the two-strain model and $S_i = N_i - \sum_{j=0}^m (\kappa_j I_{i-j})$ for the one-strain model. Further details on this method are described in Supplementary Information.

Issues pertinent to the cholera data set. Case data include all Classical and El Tor cholera patients from the Matlab surveillance area. After 1978, this area consisted of a maternal, child health and family planning treatment area and a comparison area. Cholera treatment and dynamics did not differ between these areas; we therefore aggregated these two areas in our analyses. Over the period 1966–2003, four vaccine trials were conducted in the Matlab area. To determine whether these trials affected the results, the time series were adjusted (with supplemented cases) to eliminate the protective effect of the vaccines. A fit of the model to these adjusted time series showed no appreciable difference in results (not shown). Implicit in the fit of this nonlinear disease model is the assumption that the generation time of the disease is about 1 month. Although individuals are usually symptomatic for less than a month, infected individuals can shed the bacterium for up to 3 weeks. Further analyses using data aggregated twice a month generate similar results.

Isolation of frequency components from transmission rates and climate data. Eigendecomposition analysis was applied to the climate time series. Eigendecomposition partitions signals on the basis of adaptive nonparametric functions²⁶. This technique avoids the bias towards the concentration of noise in certain frequency components. A covariance matrix is first constructed whose entry in row i , column j is the covariance of the data at lag $i - j$. The order of the eigendecomposition (that is, the embedding dimension) corresponds to the number of rows in the matrix. Singular value decomposition is then applied to this matrix to extract the eigenvectors and their corresponding eigenvalues. The eigenvalues quantify the variance associated with each eigenvector; they were normalized to have their relative contributions rescaled to 100%. We applied a decomposition of 40th order, which provided a good signal-to-noise separation, and verified that results were insensitive to parameter modifications. Projection of the signal onto the eigenmodes gives the principal components. To reconstruct the time series, the principal components associated with the first group of significant eigenvalues are combined. Six eigenmodes were kept here because they maximized signal-to-noise ratios. Residuals passed all white noise tests. The resulting reconstructed component was filtered with wavelet analysis to separate the variance into two components, one with periods more than 7 years and the other with periods between 1 and 7 years.

Global and local correlation coefficients. To determine the significance of global linear correlations between two time series that share a similar range of preselected periods, we used a bootstrap method that generates 10,000 surrogate data sets for one of the time series by randomizing the phases while preserving the power spectrum and autocorrelation function²⁷. Previous work has shown that at ENSO timescales, associations between cholera and climate variability can be discontinuous in time¹⁷. We therefore also used a time series method (scale-dependent correlation analysis) that specifically quantifies the strength of the association between variables locally in time by using correlation coefficients computed within short, truncated time windows. See the legend to Supplementary Fig. S1 for further details and an example.

Received 1 April; accepted 16 May 2005.

1. Bouma, M. J. & Pascual, M. Seasonal and interannual cycles of endemic cholera in Bengal 1891–1940 in relation to climate and geography. *Hydrobiologia* **460**, 147–156 (2001).
2. Hay, S. I. et al. Etiology of interepidemic periods of mosquito-borne disease. *Proc. Natl Acad. Sci. USA* **97**, 9335–9339 (2000).
3. Taubes, G. Global warming: Apocalypse Not. *Science* **278**, 1004–1006 (1997).

4. Rogers, D. J. *et al.* Satellite imagery in the study and forecast of malaria. *Nature* **415**, 710–715 (2002).
5. Anderson, R. M. & May, R. M. *Infectious Diseases of Humans: Dynamics and Control* (Oxford Univ. Press, Oxford, 1991).
6. Schwartz, I. B. Small amplitude, long period outbreaks in seasonally driven epidemics. *J. Math. Biol.* **30**, 473–491 (1992).
7. Zhou, G. *et al.* Association between climate variability and malaria epidemics in the East African highlands. *Proc. Natl Acad. Sci. USA* **101**, 2375–2380 (2004).
8. Pascual, M. *et al.* Cholera dynamics and El Niño–Southern Oscillation. *Science* **289**, 1766–1769 (2000).
9. Glass, R. I. *et al.* Endemic cholera in rural Bangladesh, 1966–1980. *Am. J. Epidemiol.* **116**, 959–970 (1982).
10. Koelle, K. & Pascual, M. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *Am. Nat.* **163**, 901–913 (2004).
11. Lui, W. M., Hethcote, H. W. & Levin, S. A. Dynamical behaviour of epidemiological models with nonlinear incidence rates. *J. Math. Biol.* **25**, 359–380 (1987).
12. de Jong, M. C. M., Diekmann, O. & Heesterbeek, H. in *Epidemic Models: Their Structure and Relation to Data* (ed. Mollison, D.) 84–94 (Cambridge Univ. Press, Cambridge, 1995).
13. Hastie, T. J. & Tibshirani, R. J. *Generalized Additive Models. Monographs on Statistics and Applied Probability* (Chapman & Hall, London, 1990).
14. Levine, M. M. & Pierce, N. F. in *Cholera* (eds Barua, D. & Greenough, W. B.) 285–327 (Plenum Medical, New York, 1992).
15. Clemens, J. D. *et al.* Biotype as determinant of natural immunising effect of cholera. *Lancet* **337**, 883–884 (1991).
16. Pascual, M., Bouma, M. & Dobson, A. P. Cholera and climate: revisiting the quantitative evidence. *Microbes Infect.* **4**, 237–245 (2002).
17. Rodó, X. *et al.* ENSO and cholera: A nonstationary link related to climate change? *Proc. Natl Acad. Sci. USA* **99**, 12901–12906 (2002).
18. Parthasarathy, B., Munot, A. A. & Kothwale, D. R. *Monthly and Seasonal Rainfall Series for all-India Homogeneous Regions and Meteorological Subdivisions: 1871–1994* (Research Report RR-065, Indian Institute of Tropical Meteorology, Pune, 1995).
19. Siddique, A. K. *et al.* Survival of classic cholera in Bangladesh. *Lancet* **337**, 1125–1127 (1991).
20. Kawamura, R. A possible mechanism of the Asian summer monsoon–ENSO coupling. *J. Met. Soc. Jpn* **76**, 1027 (1998).
21. Shukla, J. & Fennessy, M. J. *Simulation and Predictability of Monsoons* 567–575 (Proc. Int. Conf. on Monsoon Variability and Prediction, Technical Report WCRP-84, World Climate Research Programme, Geneva, 1994).
22. Lau, N.-C. & Nath, M. J. Impact of ENSO on the variability of the Asian–Australian monsoons as simulated in GCM experiments. *J. Clim.* **13**, 4287–4309 (2000).
23. Lau, N.-C. & Nath, M. J. Coupled GCM simulation of atmosphere–ocean variability associated with zonally asymmetric SST changes in the tropical Indian Ocean. *J. Clim.* **17**, 245–265 (2004).
24. Webster, P. J. *et al.* in *Meteorology at the Millennium. Int. Geophys. Series Vol. 83*, 198–219 (Academic, New York, 2002).
25. Dixon, P. A., Milicich, M. J. & Sugihara, G. Episodic fluctuations in larval supply. *Science* **283**, 1528–1530 (1999).
26. Elsner, J. B. & Tsonis, A. A. *Singular Spectrum Analysis: A New Tool in Time Series Analysis* (Plenum, New York, 1996).
27. Kaplan, D. & Glass, L. *Understanding Nonlinear Dynamics. Text in Applied Mathematics* (eds Marsden, J. E. *et al.*) 342–346 (Springer, New York, 1995).
28. Waldor, M. K., Colwell, R. & Mekalanos, J. J. The *Vibrio cholerae* O139 serogroup antigen includes an O-antigen capsule and lipopolysaccharide virulence determinants. *Proc. Natl Acad. Sci. USA* **91**, 11388–11392 (1994).
29. Glass, R. I. & Black, R. E. in *Cholera* (eds Barua, D. & Greenough, W. B.) 129–154 (Plenum Medical, New York, 1992).
30. Woodward, W. E. & Mosley, W. H. The spectrum of cholera in rural Bangladesh. II. Comparison of El Tor Ogawa and Classical Inaba infection. *Am. J. Epidemiol.* **96**, 342–351 (1972).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Sack for discussions on cholera immunity, K. Streatfield for support with the cholera data, A. Dobson for comments on the manuscript, P. Webster and the Climate Forecast Applications Project at Georgia Tech for the river discharge data, and the Bangladesh Water Development Board, Dhaka, Bangladesh, for flood area data. M.P. acknowledges the joint support of the NSF-NIH (Ecology of Infectious Diseases) and NOAA (Oceans and Health), as well as funding from NOAA's Joint Program on Climate Variability and Human Health, with EPRI-NSF-NASA, under which the work was initiated. Further support was provided by the James S. McDonnell Foundation Centennial fellowship to M.P. and by ICREA and an AGAUR-DURSI grant to X.R.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to M.P. (pascual@umich.edu).

Genetic interactions between polymorphisms that affect gene expression in yeast

Rachel B. Brem^{1*}, John D. Storey^{2*}, Jacqueline Whittle³ & Leonid Kruglyak⁴

Interactions between polymorphisms at different quantitative trait loci (QTLs) are thought to contribute to the genetics of many traits, and can markedly affect the power of genetic studies to detect QTLs¹. Interacting loci have been identified in many organisms^{1–5}. However, the prevalence of interactions^{6–8}, and the nucleotide changes underlying them^{9,10}, are largely unknown. Here we search for naturally occurring genetic interactions in a large set of quantitative phenotypes—the levels of all transcripts in a cross between two strains of *Saccharomyces cerevisiae*⁷. For each transcript, we searched for secondary loci interacting with primary QTLs detected by their individual effects. Such locus pairs were estimated to be involved in the inheritance of 57% of transcripts; statistically significant pairs were identified for 225 transcripts. Among these, 67% of secondary loci had individual effects too small to be significant in a genome-wide scan. Engineered polymorphisms in isogenic strains confirmed an interaction between the mating-type locus *MAT* and the pheromone response gene *GPA1*. Our results indicate that genetic interactions are widespread in the genetics of transcript levels, and that many QTLs will be missed by single-locus tests but can be detected by two-stage tests that allow for interactions.

Most heritable traits are affected by the inheritance of alleles at multiple loci, and the identification of these loci is a key challenge of modern genetic research. We recently showed that gene expression levels in yeast typically show multigenic inheritance and provide a good model for investigating the genetic basis of complex traits^{7,11,12}. Here we use this system to examine the prevalence of genetic interactions in a large set of phenotypes. A genetic interaction between a pair of loci (sometimes termed epistasis) occurs when the effect of an allele at one locus changes as a function of the allele at the other. Previous biometric analyses have provided evidence for many interactions underlying transcript levels^{6,7}. We sought to identify the loci involved in such interactions in a cross between a laboratory strain of yeast, BY, and a wild strain, RM. For this analysis we used previously described genotype and gene expression data from 112 segregants⁷. We first tested all possible pairs of loci for evidence of interaction underlying each transcript level. This strategy had little statistical power to map interactions, because of its requirement for a very large number of tests and the corresponding stringent correction of significance thresholds for multiple testing¹³.

To improve power we used a two-stage search strategy. For each transcript, we first identified the 'primary' QTL with the strongest individual linkage. We then partitioned segregants on the basis of inheritance (either BY or RM) at the primary locus and tested each subgroup for further 'secondary' loci. We computed the joint significance of the strongest such secondary QTL with the primary locus by using a newly developed statistical method based on

estimation of the false discovery rate¹³ (see Methods). This method provides an estimate of the overall fraction of transcripts for which both loci are involved in the genetics, as well as a probability for each transcript that both loci are true positives. This analysis indicated that the locus pairs identified for 57% of all transcripts were true positives, and it identified locus pairs for 225 transcripts at a false discovery rate of 5% (that is, fewer than 12 of these are expected to have either locus as a false positive).

Because the two-stage search takes account of inheritance at the primary locus in assessing the effect of the secondary locus, it can identify both locus pairs with interactions and locus pairs that act additively. We tested each of the 225 locus pairs for interacting effects by fitting a regression model relating inheritance at the linking loci to the transcript level. The model included an additive effect for each locus and a term representing the interaction between loci; the latter term was significant at $P < 0.05$ for 65% of transcripts. For comparison, we considered 547 transcripts for which two loci were identified as significant by their individual effects. Of these, only 13% had an interaction term significant at $P < 0.05$. A false discovery rate analysis¹⁴ indicated that the interaction term represented a true positive for 91% of locus pairs mapped by the two-stage search, compared with 13% of pairs mapped independently.

We next sought to determine whether the secondary loci mapped by the two-stage search could have been detected by a single-locus linkage test. Under some interaction models, the individual effects of the loci are undetectable, whereas under others the loci retain an individual effect. We tested the individual effect of each secondary locus on the corresponding transcript and found that 87% of these were significant at $P < 0.05$, indicating that most secondary loci do have a residual individual effect. However, this significance threshold does not take into account the multiple tests performed in a genome scan. When we examined how many of the secondary loci were detectable at a false discovery rate of 5% in the context of a genome scan⁷, we found that only 74 (33%) of the 225 were, indicating that 67% of secondary loci would have been missed without the two-stage search.

To identify pairs of interacting QTLs with effects on many transcripts, we constructed a histogram of the genetic positions of linking locus pairs, analogous to previous analyses with single loci^{11,15}. Most of the pairs affected single transcripts, but a few affected multiple transcripts (Fig. 1). The largest number of transcripts linking to a single locus pair was 14, with four additional transcripts linking nearby (Fig. 1); because this group contained both YCR040W/*MAT* α 1 and its silenced copy YCL066W/*HML* α 1, we eliminated the latter from further analysis, leaving 17 transcripts that linked to this locus pair. The primary QTL of the pair lay near *MAT*, which confers α or *a* mating type on a haploid yeast cell

¹Program in Computational Biology, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N, M2-B876, Seattle, Washington 98109, USA. ²Department of Biostatistics and Department of Genome Sciences, University of Washington, Campus Mail Stop 357232, Seattle, Washington 98195-7232, USA. ³Infectious Disease Research Institute, 1124 Columbia Street, Suite 600, Seattle, Washington 98104, USA. ⁴Lewis-Sigler Institute for Integrative Genomics and Department of Ecology and Evolutionary Biology, Carl Icahn Laboratory, Princeton University, Princeton, New Jersey 08544-1003, USA.

*These authors contributed equally to this work.

depending on the integration of genes at the locus. The secondary locus contained the gene encoding the G-protein subunit *Gpa1*, in which a single polymorphism in the BY parent strain, S469I, has previously been shown to affect the expression of pheromone response genes¹². Of the 17 transcripts whose linkage to the *MAT* and *GPA1* loci was detected by the two-stage search, 11 were previously shown to be regulated by mating type (of which seven are also regulated by the pheromone response pathway); of the remaining six linking transcripts, two have a known function that does not involve mating, and four have no known function¹⁶. For 14 of the transcripts, the *MAT* and *GPA1* loci showed direct evidence for interaction in the regression model at $P < 0.05$; the independent effect of the *GPA1* locus exceeded the genome-wide significance threshold for single-locus linkage for only one of these 17 transcripts.

To test the hypothesis that variation at *MAT* and the S469I mutation in *GPA1* interact genetically, we engineered isogenic yeast strains carrying each of the four possible combinations of alleles at the two loci. For each of these combinations we compared expression in the engineered strains with expression in segregants having the same allele combination (but having varying inheritance for the rest of the genome). For 7 of 11 transcripts previously known to be regulated by mating type, and 3 of 4 transcripts of unknown function, the pattern of effect of the *MAT*–*GPA1* genotype on expression was the same in engineered strains as in segregants (Fig. 2 and Supplementary Information), indicating that variation at *MAT* and *GPA1* is sufficient to explain transcript level differences across segregants. YOR090C/*PTC1* and YOR162C/*YRR1*, which have known functions unrelated to mating, did not show agreement between engineered strains and segregants (Supplementary Information), indicating that these effects might represent either false positives or QTLs linked, but unrelated, to *MAT* or *GPA1*.

In a wild-type haploid cell exposed to pheromone, *GPA1* activates the pheromone response pathway to prepare for mating^{17,18}. We showed previously that, in the absence of endogenous pheromone, the S469I variant is associated with the upregulation of genes involved in mating-type-independent mating functions (for example cytoskeletal rearrangement and cell fusion)¹². By contrast, most genes affected by the genetic interaction between *MAT* and *GPA1* perform mating-type-specific mating functions (for example pheromone detection and cell conjugation). Levels of α -specific transcripts in the **a** background have been measured at less than one copy per cell¹⁹. This indicates a possible molecular model for the interaction between *MAT* and *GPA1*: because of tight repression by *MAT*, other regulators have no impact on α -specific transcripts in the α background, and

vice versa (Fig. 2a–d and Supplementary Information). This situation is analogous to a polymorphism with effects on sex-limited traits in each of the two sexes.

GPA1 polymorphism affected the expression of *MAT* α 1 and *MAT* α 2 (Fig. 2e, f). This is in contrast to results from genome-wide studies^{18,20} and a classic molecular biology report²¹ showing no significant regulation of *MAT* genes by the pheromone response pathway. The discrepancy could indicate that the S469I variant of *Gpa1* affects regulation directly at *MAT* through a different pathway from that of exogenous pheromone. Alternatively, *GPA1* might act indirectly through other regulators of *MAT*²². We speculate that expression changes due to polymorphism at *GPA1* might affect the regulatory activity of *MAT*-encoded proteins on their downstream targets. For several transcripts, polymorphism in *GPA1* seemed to show opposing effects in the two mating types (Fig. 2e, f and Supplementary Information); whether this represents a true result or an artefact is unclear.

Mapping of genetic factors that underlie polygenic physiological, medical and agricultural traits in outbred populations has met with limited success. It has been suggested that this might be mainly because the sizes of individual locus effects are modest, owing to genetic interactions between loci¹, but the prevalence of interactions has not been well characterized. Our results indicate that genetic interactions underlie the inheritance of roughly half of all transcript levels in yeast and that at least one member of an interacting locus pair typically has too small an individual effect to be identified on its own. We have shown that a two-stage search provides a useful strategy for the identification of such secondary loci. Because this

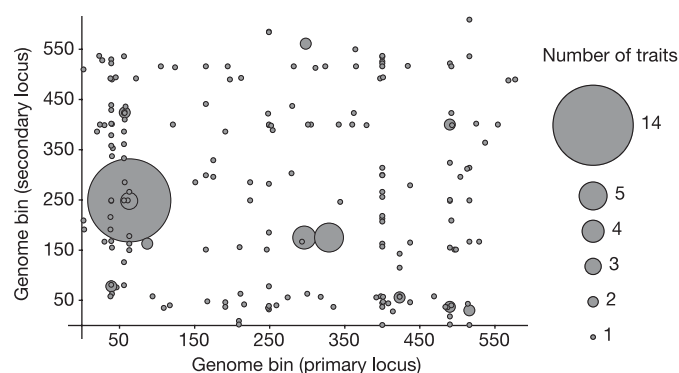


Figure 1 | Genome distribution of QTL pairs detected by the two-stage linkage search. On each axis the genome is divided into 611 bins of 20 kilobases each, shown in chromosomal order. The set of transcripts mapping to QTLs in each pair of bins in the two-stage analysis is represented as a circle, with the width proportional to the number of such co-linking transcripts (a scale is shown at the right); circles are centred on the corresponding bins. The largest circle represents the *MAT*–*GPA1* locus pair.

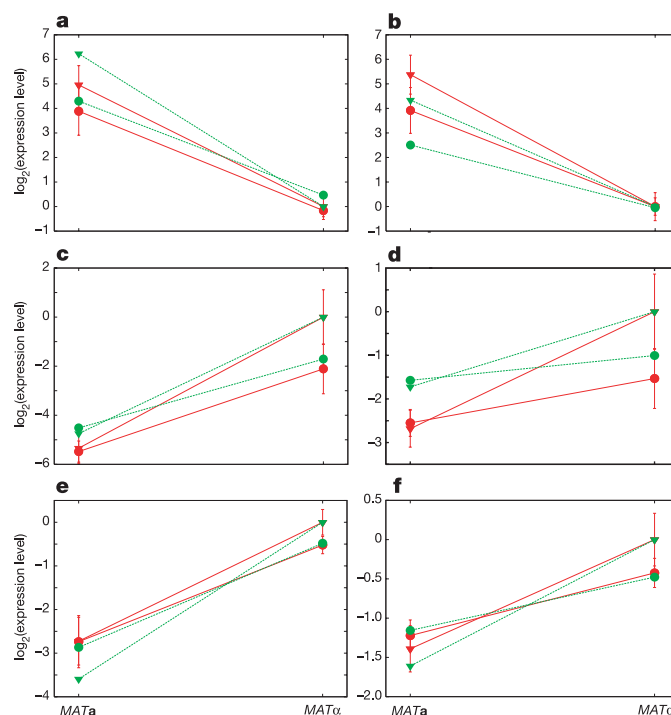


Figure 2 | Example transcripts showing genetic interaction between *MAT* and *GPA1*. Each panel represents one transcript: **a**, YFL026W/*STE2*; **b**, YIL015W/*BAR1*; **c**, YJR004C/*SAG1*; **d**, YGL090W/*LIF1*; **e**, YCR040W/*MAT* α 1; **f**, YCR039C/*MAT* α 2. Each red point represents the mean expression level over segregants with one genotype, normalized by the mean over segregants inheriting the BY allele at both loci; red error bars represent standard deviations among segregants with the respective genotype. Each green point represents the expression level in one engineered strain, normalized by the level in the engineered strain with the BY allele at both loci. Strain mating type is indicated on the x axis (*MAT*_a is the RM allele, *MAT*_α is the BY allele). Triangles and circles represent strains carrying the BY and RM alleles of *GPA1*, respectively.

strategy relies on the identification of primary loci by their individual effects, the detection of interacting locus pairs in which both individual effects are small remains a challenge, so the overall prevalence of interactions may be even higher than estimated here. The discovery here of previously uncharacterized genetic effects in the well-studied yeast pheromone response network underscores the importance of interaction mapping in genetic analyses.

METHODS

Strains and expression measurements. Segregants, genotypes, expression measurements and single-locus linkage results were those of ref. 7. To find multiple independent linkages, for each transcript we identified the marker with the strongest linkage score on each chromosome; if two or more such chromosome peaks exceeded the genome-wide single-locus cut-off in ref. 7, we classified these as multiple independent QTLs. For the direct test of *MAT-GPA1* interaction, the I469 and S469 alleles of *GPA1* were each engineered into the S288c derivatives BY4709 (*MAT α URA3 Δ 0*) and BY4724 (*MAT α LYS2 Δ 0 URA3 Δ 0*) as in ref. 12; expression arrays were performed as in ref. 7, except that the reference sample was a 1:1 mixture of RNAs from the BY and RM strains. For two-stage mapping and interaction tests on the resulting loci, spatial loess normalization was performed on expression data using the genes in ref. 12. Missing genotype data were imputed with the use of a standard hidden Markov model algorithm implemented in R/qtl (ref. 23). Missing expression data were imputed by using a $K = 15$ nearest-neighbours method²⁴.

Two-locus mapping. A two-stage procedure was employed to identify pairs of linked loci for each expression trait. For each transcript and marker, a Wilcoxon rank-sum statistic was formed to quantify expression differences between the segregants grouped by inheritance at the locus. We identified the 'primary' QTL for each transcript as the locus with the most significant Wilcoxon rank-sum statistic. We then partitioned segregants on the basis of inheritance (either BY or RM) at the primary locus and similarly tested each subgroup for further 'secondary' loci. The locus with the highest statistic between either partition was chosen as the secondary locus. At both stages, the expression traits were randomly permuted (five times) and analogous maximal statistics were recalculated from this null distribution. These null statistics were pooled across transcripts, for a total of about 30,000 at each stage.

We considered a transcript to be a 'false discovery' if either the primary or the secondary locus was a false positive. Under this definition of a false discovery, it is not straightforward to calculate a P value for each transcript because the null distribution must account for an unknown mixture of three situations: both loci are false positives, the primary locus only is a false positive, or the secondary locus only is a false positive. We therefore employed a new statistical method to rank the expression traits for significance and calculate the false discovery rate for each significance cut-off²⁵. In brief, we formed nonparametric empirical Bayes estimates of the posterior probability that the primary QTL is a true positive, and then the posterior probability that the secondary QTL is a true positive given that the primary locus is a true positive. The joint probability that both loci are true positives is equal to the product of these two probabilities. The observed and permutation-based null statistics for each stage were used to form conservative estimates of these probabilities (ref. 13). In the Bayesian setting, the false discovery rate is exactly equal to the probability that a transcript is a false positive given that it is called significant²⁵. Equivalently, it can be written as one minus the probability that the transcript is a true positive, given that it is called significant. The estimated joint linkage probabilities can be used to estimate this latter quantity directly, yielding an estimate of the false discovery rate for any chosen joint linkage probability cut-off. With a cut-off corresponding to a 5% false discovery rate, 225 expression traits were called significant for two-locus linkage with this method.

Interaction test. For each locus pair mapped in the two-stage model or mapped independently, we fitted the model

$$t = ax + by + cxy + d$$

over all segregants. Here t is $\log_2$ of the ratio of expression between the strain of interest and a reference sample, a , b , c , and d are parameters that are specific to the given transcript, x is inheritance at the first locus, and y is inheritance at the second locus. A standard F -test was used to test the null hypothesis that $c = 0$ by comparing the goodness-of-fit of the above full model with that of the purely additive model $t = ax + by + d$. Because these tests were performed on transcripts and locus pairs previously mapped by the two-stage or independent linkage calculations, there is a potential for the significance to be artificially inflated. However, because the model used in each linkage search is a restricted version of the purely additive model, the interaction term can in fact be tested on previously mapped locus pairs without incurring a bias. The program QVALUE

(faculty.washington.edu/~jstorey/qvalue/) was applied to these P values to estimate the total proportion of transcripts showing evidence for interaction.

Received 8 March; accepted 2 June 2005.

- Carlberg, O. & Haley, C. S. Epistasis: too often neglected in complex trait studies? *Nature Rev. Genet.* **5**, 618–625 (2004).
- Leamy, L. J., Workman, M. S., Routman, E. J. & Cheverud, J. M. An epistatic genetic basis for fluctuating asymmetry of tooth size and shape in mice. *Heredity* **94**, 316–325 (2005).
- Montooth, K. L., Marden, J. H. & Clark, A. G. Mapping determinants of variation in energy metabolism, respiration and flight in *Drosophila*. *Genetics* **165**, 623–635 (2003).
- Shook, D. R. & Johnson, T. E. Quantitative trait loci affecting survival and fertility-related traits in *Caenorhabditis elegans* show genotype–environment interactions, pleiotropy and epistasis. *Genetics* **153**, 1233–1243 (1999).
- Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer, Sunderland, Massachusetts, 1998).
- Gibson, G. *et al.* Extensive sex-specific nonadditivity of gene expression in *Drosophila melanogaster*. *Genetics* **167**, 1791–1799 (2004).
- Brem, R. B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl Acad. Sci. USA* **102**, 1572–1577 (2005).
- Kroymann, J. & Mitchell-Olds, T. Epistasis and balanced polymorphism influencing complex trait variation. *Nature* **435**, 95–98 (2005).
- Rawson, P. D. & Burton, R. S. Functional coadaptation between cytochrome *c* and cytochrome *c* oxidase within allopatric populations of a marine copepod. *Proc. Natl Acad. Sci. USA* **99**, 12955–12958 (2002).
- Caicedo, A. L., Stinchcombe, J. R., Olsen, K. M., Schmitt, J. & Purugganan, M. D. Epistatic interaction between *Arabidopsis* *FRI* and *FLC* flowering time genes generates a latitudinal cline in a life history trait. *Proc. Natl Acad. Sci. USA* **101**, 15670–15675 (2004).
- Brem, R. B., Yvert, G., Clinton, R. & Kruglyak, L. Genetic dissection of transcriptional regulation in budding yeast. *Science* **296**, 752–755 (2002).
- Yvert, G. *et al.* Trans-acting regulatory variation in *Saccharomyces cerevisiae* and the role of transcription factors. *Nature Genet.* **35**, 57–64 (2003).
- Storey, J. D., Akey, J. M. & Kruglyak, L. Multiple locus linkage analysis of genome-wide expression in yeast. *PLoS Biol.* **3**, e267 (2005).
- Storey, J. D. A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 479–498 (2002).
- Schadt, E. E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
- Csank, C. *et al.* Three yeast proteome databases: YPD, PombePD, and CalPD (MycopathPD). *Methods Enzymol.* **350**, 347–373 (2002).
- Bardwell, L., Cook, J. G., Inouye, C. J. & Thorner, J. Signal propagation and regulation in the mating pheromone response pathway of the yeast *Saccharomyces cerevisiae*. *Dev. Biol.* **166**, 363–379 (1994).
- Zeitlinger, J. *et al.* Program-specific distribution of a transcription factor dependent on partner transcription factor and MAPK signalling. *Cell* **113**, 395–404 (2003).
- Velculescu, V. E. *et al.* Characterization of the yeast transcriptome. *Cell* **88**, 243–251 (1997).
- Roberts, C. J. *et al.* Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**, 873–880 (2000).
- Hagen, D. C. & Sprague, G. F. Jr Induction of the yeast α -specific STE3 gene by the peptide pheromone α -factor. *J. Mol. Biol.* **178**, 835–852 (1984).
- Sudarsanam, P., Iyer, V. R., Brown, P. O. & Winston, F. Whole-genome expression analysis of *snf/swi* mutants of *Saccharomyces cerevisiae*. *Proc. Natl Acad. Sci. USA* **97**, 3364–3369 (2000).
- Broman, K. W., Wu, H., Sen, S. & Churchill, G. A. R/qtl: QTL mapping in experimental crosses. *Bioinformatics* **19**, 889–890 (2003).
- Troyanskaya, O. *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520–525 (2001).
- Storey, J. D. The positive false discovery rate: A Bayesian interpretation and the q -value. *Ann. Stat.* **31**, 2013–2035 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Botstein and J. Broach for reading the manuscript and for discussions, E. Smith for constructing plasmids, and E. Foss for providing strains. The experiments were performed when J.W. and L.K. were at the Fred Hutchinson Cancer Research Center and the Howard Hughes Medical Institute. This work was supported by funding from the Howard Hughes Medical Institute (to L.K.) and grants from the National Institutes of Health (to L.K. and J.D.S.). L.K. is a James S. McDonnell Centennial Fellow. R.B. is supported by a Burroughs-Wellcome Career Award at the Scientific Interface.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.B.B. (rbrem@fhcrc.org) or L.K. (leonid@genomics.princeton.edu).

LETTERS

Centrosome localization determines neuronal polarity

Froylan Calderon de Anda^{1*}, Giulia Pollarolo^{1*}, Jorge Santos Da Silva^{1*}, Paola G. Camoletto¹, Fabian Feiguin¹ & Carlos G. Dotti^{1,2}

Neuronal polarization occurs shortly after mitosis. In neurons differentiating *in vitro*, axon formation follows the segregation of growth-promoting activities to only one of the multiple neurites that form after mitosis^{1,2}. It is unresolved whether such spatial restriction makes use of an intrinsic program, like during *C. elegans* embryo polarization³, or is extrinsic and cue-mediated, as in migratory cells⁴. Here we show that in hippocampal neurons *in vitro*, the axon consistently arises from the neurite that develops first after mitosis. Centrosomes, the Golgi apparatus and endosomes cluster together close to the area where the first neurite will form, which is in turn opposite from the plane of the last mitotic division. We show that the polarized activities of these organelles are necessary and sufficient for neuronal polarization: (1) polarized microtubule polymerization and membrane transport precedes first neurite formation, (2) neurons with more than one centrosome sprout more than one axon and (3) suppression of centrosome-mediated functions precludes polarization. We conclude that asymmetric centrosome-mediated dynamics in the early post-mitotic stage instruct neuronal polarity, implying that pre-mitotic mechanisms with a role in division orientation may in turn participate in this event.

To identify when neuronal polarity is first determined, we analysed the morphological differentiation of recent post-mitotic hippocampal neurons in culture⁵. Instead of the standard 18-day-old rat embryos, hippocampi were taken from 16-day-old embryos, to maximize the number of post-mitotic cells. Shortly after attachment to the extracellular matrix (laminin or poly-L-lysine), numerous cells present a smooth cell body (Fig. 1a). In these, the first sign of morphological differentiation is the emergence of a lamellipodium restricted to a small area of the cell surface, followed by the sequential appearance of further, similar extensions (Fig. 1a). Continuous observation revealed that neurites form from the different lamellipodia with identical order of appearance—that is, the first neurite forms from the first lamellipodium (Fig. 1b). After 24 h, a longer neurite can be seen from the area where the first lamellipodium appeared (Fig. 1b, rightmost panel). By 48 h, this neurite is twice as long as any of the later-appearing neurites and has acquired tau-1 reactivity, a typical feature of axons (Fig. 1c). This type of progression was frequently observed but did not occur in all neurons (see Supplementary Fig. 1a).

To assess whether polarized growth information is present as early as when the first lamellipodium/neurite arises, neurons with a single lamellipodium (as in Fig. 1a) were evaluated for their response to a low dose of the actin-depolymerizing drug cytochalasin D, which emphasizes axonal growth competence⁶. This treatment results in the sprouting of all neurites from the area of the first lamellipodium (Fig. 1d). Together, these results indicate that the area of the cell

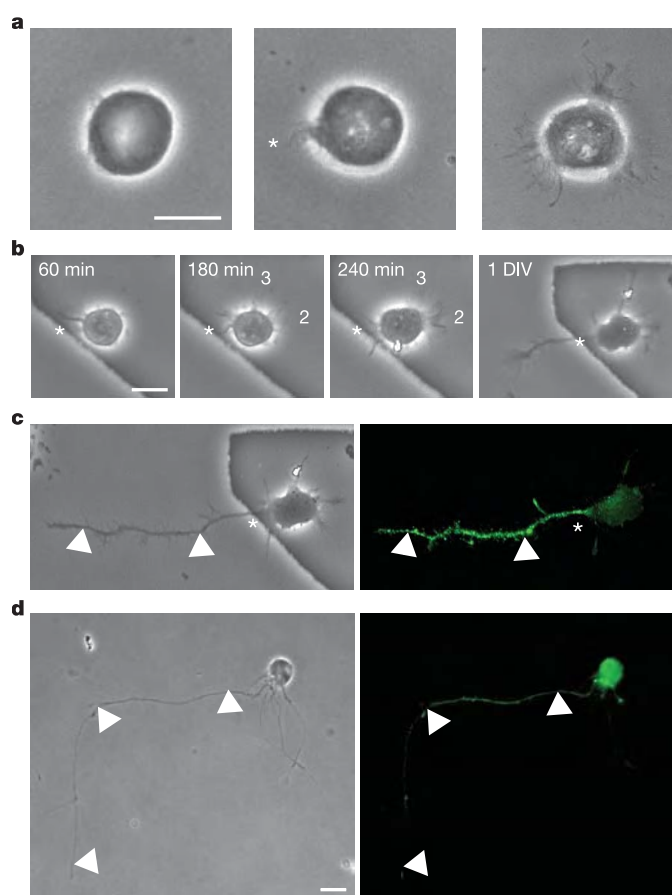


Figure 1 | The first sprout contains polarized growth information.

a, Immediately after attachment, some hippocampal neurons from 16-day-old embryos have a smooth cell body (left). A thin lamellipodium then appears (asterisk in middle panel), followed by the appearance of similar extensions (right panel). **b**, Same cell shown in the middle panel of **a**. The first lamellipodium acquires neurite shape (asterisk at 60 min) sooner than the later-appearing lamellipodia (compare asterisk with lamellipodia 2 and 3 in images at 180 and 240 min). After 24 h *in vitro*, the first neurite is significantly longer (asterisk in rightmost panel). DIV, days *in vitro*. **c**, The same cell as **b**, one day later. The neurite that formed first (asterisk) has the typical length of an axon in these cells (arrowheads in left panel), confirmed by tau-1 immunoreactivity (green, arrowheads in right panel). **d**, In cells with a single sprout (as in **a**, middle panel), cytochalasin D results in the outgrowth of all neurites from a single pole (left). Only one neurite (arrowheads in both panels) is tau-1-positive (right). All scale bars, 10 μm.

¹Cavaliere Ottolenghi Scientific Institute, Università degli Studi di Torino, 10043 Orbassano, Torino, Italy. ²Department of Human Genetics, Catholic University of Leuven and Flanders Interuniversity Institute of Biotechnology, 3000 Leuven, Belgium.

*These authors contributed equally to this work.

where the first membrane breakage occurs contains most of the neuronal growth potential. Moreover, the initial sprout and longer neurite of neighbour neurons in the germinal layer *in situ* present the same spatial orientation (see Supplementary Fig. 1b), indicating that the correlation between first membrane breakage and morphological polarization is not an exclusive *in vitro* response.

The area of the neuron with concentrated polarity information might be defined by the polarized organization and dynamics of the cytoplasm, through centrosome-dependent microtubule nucleation as well as Golgi-endosome positioning and dynamics⁷. In fact, the

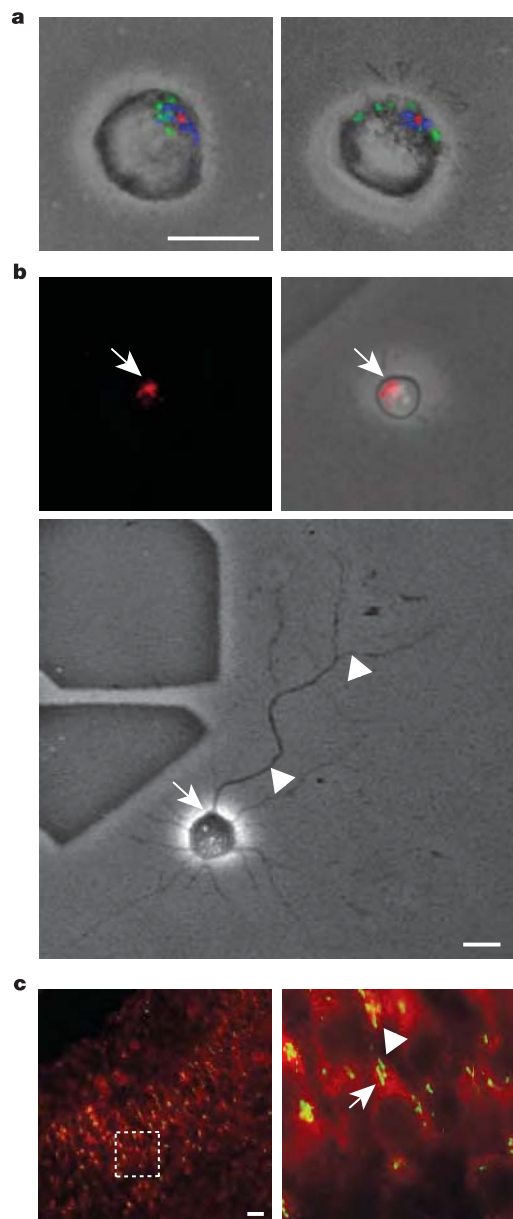


Figure 2 | Cytoplasmic asymmetry marks the area of neuronal polarization. **a**, Centrosomes (red), Golgi apparatus (blue) and endosomes (green) are polarized in round cells (left) and at the base of the first lamellipodium (right). Scale bar, 10 μ m. **b**, Labelling of the Golgi apparatus in live cells (using bodipy-ceramide) before and after polarization. Note that the pole of organelle clustering before polarization (arrowheads, top images) is that from which the axon forms (arrowheads, bottom image). Scale bar, 10 μ m. **c**, Left panel shows localization of the Golgi apparatus (green, GM130) in relation to the axons (red) in a coronal section of E18 rat hippocampus. Scale bar, 10 μ m. Right panel shows an enlarged view of the boxed area in the left panel. Golgi tubules (arrow) are evident at the base of the cell neurite (arrowhead). This correlation was evident in 86% of the cells ($n = 40$ cells, 5 sections analysed).

centrosomes, Golgi apparatus and late/recycling endosomes cluster together at one pole of completely undifferentiated neurons, or at the base of the only lamellipodium of more mature cells (Fig. 2a). Continuous observation of cells in which the location of the Golgi

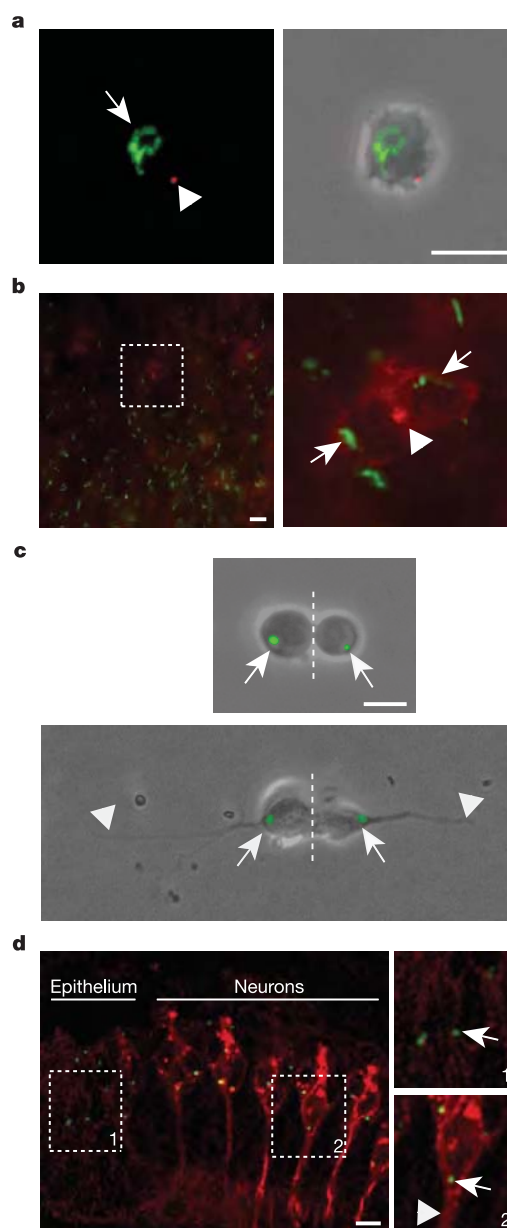


Figure 3 | Centrosome/Golgi position is opposite the plane of mitotic division. **a**, Left panel, the Golgi apparatus (marked by GM130 staining in green, arrow) is opposite the cleavage furrow (marked by citron-kinase staining in red, arrowhead) in non-differentiated hippocampal neurons. Right panel shows a phase-contrast image. **b**, Left panel, Golgi and citron-kinase staining of a coronal section of E18 rat embryo germinal layer at low magnification. Right panel shows an enlargement of the indicated area. Golgi stacks (arrows) are opposite the cleavage furrow (arrowhead). **c**, Recently generated neurons from GFP-centrosomin-expressing *Drosophila melanogaster* GMCs in culture. Top panel, centrosomes (green, arrows) are opposite to the plane of division (dotted line). Bottom panel, axons (arrowheads) arise from the centrosomal pole (arrows), opposite the plane of cleavage (dotted line). **d**, Confocal microscopy of GFP-centrosomin-expressing neurons of the third instar *Drosophila melanogaster* eye disc epithelium. Centrosomes have a basal location in epithelial neuronal precursors (box 1) and at the base of the axons of the photoreceptors (box 2). Higher magnification images of boxes 1 and 2 confirm centrosome (arrows) proximity to axons (arrowhead). Scale bar, 10 μ m (**a**, **b**), 5 μ m (**c**, **d**).

apparatus, labelled with bodipy-ceramide (see Methods), was known from the no-neurite stage, revealed that this organelle's polarization, and therefore centrosome polarization, marks the site of first neurite appearance, later becoming the axon (Fig. 2b). This type of correlation was observed in a high proportion of *in vitro* differentiating cells (Supplementary Fig. 2a), and was also evident in neurons maturing *in situ* in the germinal layer (Supplementary Fig. 2b), and also in their final position in the hippocampus (Fig. 2c).

Before addressing whether asymmetric confinement of organelles is sufficient to instruct neuronal polarization, we investigated the extent to which it relates to the plane of the previous mitotic division. This question arises because the position of centrosomes in a recent post-mitotic cell is opposite to the plane of cleavage⁸, but can change orientation during cellular maturation and also in hippocampal neurons in culture⁹. A first indication that neuronal polarization might be defined by the plane of mitotic division came from our

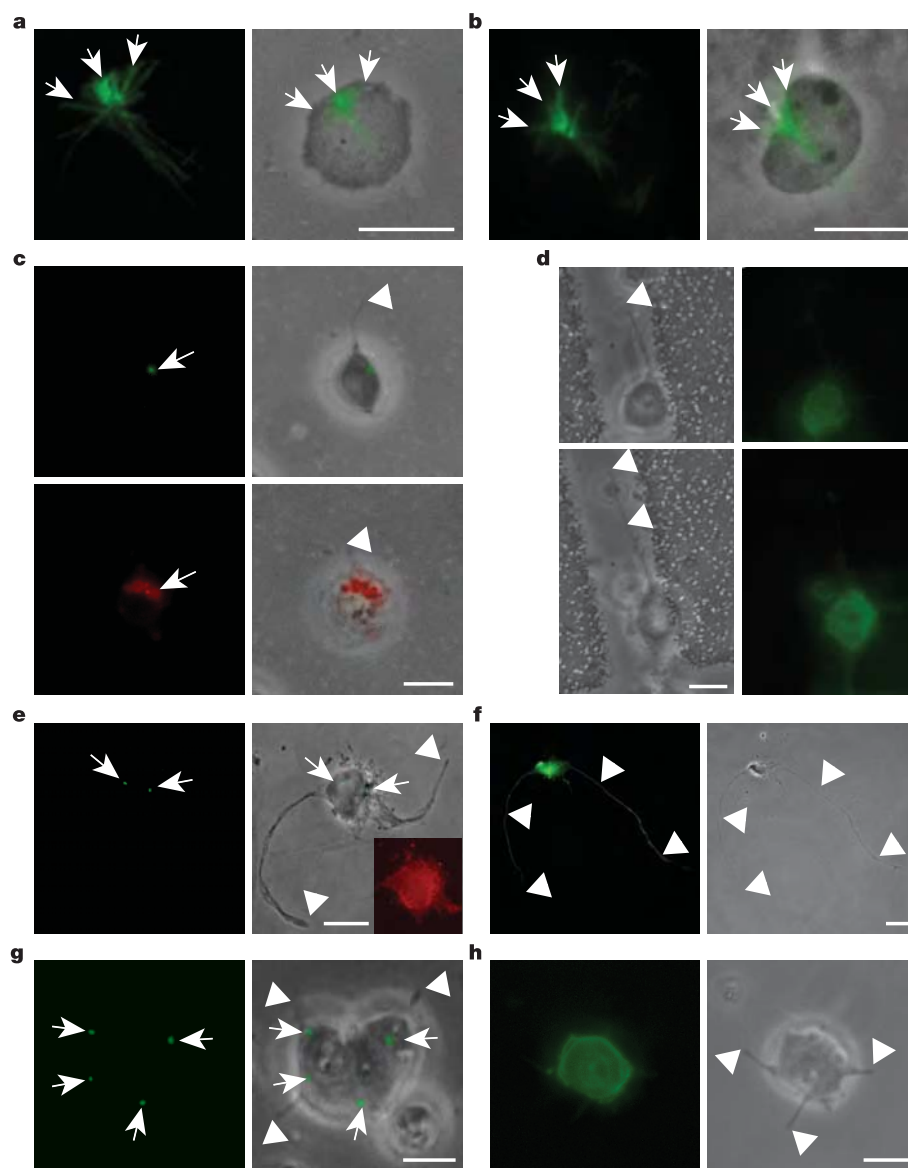


Figure 4 | Centrosomal polarized activity is necessary and sufficient to induce neuronal polarity. **a**, α -tubulin staining (green) of nocodazole-treated, recovered undifferentiated cells (see Methods). Note that microtubules present a polarized organization, with more oriented towards the nearby plasma membrane (arrows). The corresponding phase-contrast image is shown on the right. **b**, Cells were treated as in **a** and stained for acetylated microtubules (see Methods). Note that most of the stable microtubules face the nearby membrane (arrows). The corresponding phase-contrast image is shown on the right. **c**, *Drosophila*-derived third instar neurons expressing GFP-centrosomin *in vitro* before (top) and 24 h after (bottom) exposure to CALI. Top left, GFP-centrosomin before illumination (arrow). Top right, axon length before illumination (arrow). Bottom left, the treated cell retains the capacity to internalize fluid-phase ligands (arrow). Bottom right, axon retraction following CALI (compare arrowhead position in top and bottom phase-contrast images). **d**, *Drosophila*-derived third instar neurons expressing GFP-CD8 *in vitro*

before and after exposure to CALI. Top left, arrowhead marks the end of the axon before CALI. Top right, GFP-CD8 signal after CALI. Lower left, the same cell is shown 24 h after CALI. Note that the axon grew (compare arrowhead position with upper image). Lower right, GFP signal in the same cell. **e**, Cytochalasin D triggers cytokinesis arrest in embryonic hippocampal neurons. Neurons with two centrosomes (left, arrows) in a single cell body (inset in right image) sprout two neurites (arrowheads) from the ectopic centrosome-bearing hemispheres. **f**, Cytokinesis-arrested neuron 48 h after treatment. The two longest neurites (arrowheads in right panel) are tau-1-positive (arrowheads in left panel). **g**, Cytokinesis-arrested *Drosophila* neuroblasts. Ectopic axons (arrowheads in right panel) are present in the vicinity of the ectopic centrosomes (arrows). **h**, Cytochalasin-D-mediated cytokinesis arrest in *Drosophila* neuroblasts expressing the membrane protein GFP-CD8. Note that the multiple axons (arrowheads) grow from a single membrane envelope. Scale bar, 10 μ m (**a**, **b**, **e**, **f**), 5 μ m (**c**, **d**, **g**, **h**).

observation that the Golgi apparatus is found opposite to the place of last cleavage (Fig. 3a), which is identified by the presence of citron-kinase¹⁰. Although very few cells retain citron-kinase as a mitotic furrow remnant once in culture, organelle–furrow apposition was found in the majority of such cells (see Supplementary Fig. 2c). Similar organelle–cleavage furrow apposition was evident in neurons undergoing differentiation *in situ* (Fig. 3b). To determine whether the axon emerges from the pole opposite to the plane of last mitotic division, we followed the maturation process of *Drosophila melanogaster* third instar larva ganglion mother cells (GMCs) undergoing differentiation *in vitro*¹¹. In these cells, the identification of centrosome position is possible owing to the expression of green fluorescent protein (GFP)-labelled centrosomin¹². We found that in these cells, axon formation occurs opposite to the plane of mitotic cleavage, from the pole with centrosomes (Fig. 3c). This event was further corroborated by time-lapse video recordings of *Drosophila* third instar larvae brain neuroblasts undergoing division and differentiation *in vitro* (see Supplementary Fig. 3a). That the longest neurite of these cells is the axon is suggested by the enrichment in BP-102 antigen¹³ in this neurite (see Supplementary Fig. 3b). Furthermore, in the *Drosophila* third instar larvae eye disc, in which precursor epithelial cells and differentiated neurons are in the same plane of section, centrosomes are basally positioned in the epithelial cells and neurons, next to the site of axon emergence (Fig. 3d). This last result suggests that the position of centrosomes marking polarity might be defined from the pre-neuronal stage. More importantly, it confirms that centrosome localization precedes morphological polarization *in situ*.

If centrosome positioning has a role in inducing polarized growth, microtubule polymerization and membrane trafficking should be polarized towards the direction of growth. In fact, more and more packaged microtubules (Fig. 4a) with higher stability¹⁴ (Fig. 4b) are polarized towards the direction of growth (compared with the opposite or lateral sides) in the round, undifferentiated neurons. Notably, the first lamellipodium forms closest to the most abundant growing microtubules (see Supplementary Fig. 4a). Consistently, more transport of newly synthesized membrane proteins takes place towards the first lamellipodium/neurite in neurons, both *in vitro* (Supplementary Fig. 4b) and *in situ* (Supplementary Fig. 4c).

To prove that centrosome-directed polarized dynamics are required for the generation of neuronal polarity, we used pharmacological treatments to impair microtubule polymerization and Golgi–endosome membrane trafficking. Non-polarized neurons were treated with nocodazole (which prevents microtubule polymerization) or brefeldin A (which blocks protein recycling to the Golgi apparatus and endosomes) (see Methods). Not surprisingly, these treatments prevented polarization (see Supplementary Fig. 5a, b). We blocked centrosomal-mediated functions using a different approach, chromophore-assisted light inactivation (CALI)^{15,16}. This was performed in cultured *Drosophila melanogaster* neurons expressing GFP–centrosomin^{11,12}. Laser illumination of the centrosomal area in these neurons (until the GFP signal is considerably reduced) results in lack of axon formation, or lack of axonal growth in already polarized neurons (Fig. 4c). The fact that CALI does not block the internalization and retrograde transport of fluid-phase ligands (Fig. 4c) suggests that lack of polarized growth is largely the consequence of impaired centrosomal-mediated centrifugal dynamics (see Fig. 4a, b). It also suggests that CALI-induced polarization arrest is not the consequence of overt cell toxicity. Furthermore, application of CALI to neurons expressing GFP–CD8 did not perturb neuronal polarization (Fig. 4d).

Having demonstrated that centrosomal-mediated activities are required for the occurrence of neuronal polarization, we next investigated whether these activities are sufficient to induce polarization. For this, we generated neurons with more than one centrosome by blocking cytokinesis of neuroblasts present in hippocampal cultures from 16-day-old rat embryos, with a low concentration of

cytochalasin D added immediately after dissociation (see Methods and ref. 17). One day after treatment, some neurons show two centrosomal clusters within a single cell body (Fig. 4e). More importantly, instead of the normal phenotype of one axon and several short neurites, these cells have two long neurites, emerging from the poles with the ectopic centrosomes (Fig. 4e). Analysis of these cells at later times revealed that the long neurites are tau-1 positive (Fig. 4f). Although this phenotype was not very common, it was observed in all cells with two centrosomal clusters (27 cells from two different cultures). Addition of cytochalasin D to *Drosophila* neuroblasts in culture results in giant neurons containing several centrosomes within a single cell body (Fig. 4h). As in the mammalian neurons, these neurons have numerous axons, not one, emerging from the area close to the ectopic centrosomes (Fig. 4g, h).

A widely accepted concept in the field of neuronal polarity is that axonal fate is decided by a contest between the different neurites that decorate the cell body for the one-neurite-only ‘grow’ signal¹⁸. The data here presented illustrate that the instruction for polarized growth occurs before neurites are formed—in the immediate post-mitotic stage—by the presence of centrosome-associated organelle-polarized dynamics. The consequence of this intrinsic signal is a nearby membrane deformation with a predisposition for polarized growth. As any remaining deformations/neurites have the potential to become the axon⁵, the preference shown here for the first neurite might simply reflect a quantitative and/or qualitative advantage for this neurite to respond to extrinsic growth-triggering cues in a more efficient manner. This mechanism would explain why neurons over-expressing growth-promoting molecules^{19–23} can generate multiple neurites with axon-like growth. It would also explain why, when the axon of these cells is cut, the new axon forms from the longest neurite²⁴. Moreover, it provides an explanation for the observation that spatial segregation of a membrane protein to a single growth cone links the intracellular growth machinery with extracellular growth-promoting signals²⁵. Considering that the position of the centrosome organelles in the immediate post-mitotic neuron might have been instructed at the previous stage, it is reasonable to hypothesize that neuronal polarity instruction can be defined even before neurons are made. The data presented here lay the groundwork for addressing this and other intriguing mechanistic questions.

METHODS

Antibodies. The following antibodies were used: mouse monoclonal anti-tau-1 (MAB3420, Chemicon), mouse monoclonal anti- α -tubulin (CP06, Oncogene), mouse monoclonal anti-GM130 (BD Transduction Labs), mouse monoclonal anti- γ -tubulin (Sigma), rabbit polyclonal anti-citron-kinase (a gift from F. Di Cunto), rabbit polyclonal anti-HRP (ICN Biomedicals), mouse monoclonal BP 102 anti-CNS axons (developed by C. Goodman, Developmental Studies Hybridoma Bank), mouse monoclonal anti-acetylated tubulin (gift from G. Piperno²⁶) and sheep polyclonal anti-human TGN 46 (AHP500G, Serotec). Goat anti-rabbit and goat anti-mouse Alexa-fluoro 350/488/568-labelled antibodies (Molecular Probes) were used as secondary antibodies. Filamentous actin was detected with TRIC-labelled phalloidin (Sigma).

Primary cultures and pharmacological treatments. Primary cultures of rat embryo hippocampal neurons were prepared from 16-day-old embryos as described²⁷. Primary cultures from third instar *Drosophila* larval brains were prepared as described¹¹. The following stocks were used: wild-type Oregon-R; elavGal4,w; yw;tubulin Gal4/TM3Sb; w;UASp-GFP–Cnn1. Primary cultures of GFP-tagged centrosomin-expressing neurons were prepared from tubulin- or elav-driven Gal4 flies crossed with UASp-GFP–Cnn1, and the GFP-expressing larvae were identified under a fluorescence microscope as for the wild-type cultures. For cytokinesis-arrest studies, cells from 16-day-old rat hippocampal neurons or from third instar *Drosophila* larvae brain (expressing GFP–centrosomin or GFP–CD8) were dissected and treated with cytochalasin D (Sigma, 200 ng ml^{−1}) while in suspension, as described¹⁷. Nocodazole (7 μ M, Sigma) was used for microtubule depolymerization–repolymerization studies as previously published⁹. Cytochalasin D (2 μ g ml^{−1}) was used for analysis of actin depolymerization as published⁶. For perturbation of Golgi–endosomal membrane dynamics, brefeldin A (Sigma) was used at 10 μ g ml^{−1} for 5 h.

Microscopy. Immunofluorescence analysis of protein expression and distribution in cultured hippocampal neurons was done as in ref. 27. Immunofluorescence analysis of dissociated and cultured third instar *Drosophila* neurons was performed as in ref. 28. Simultaneous detection of two mouse monoclonal antibodies was performed using a Zenon–Alexa fluoro 568 mouse IgG1 kit (Molecular Probes). Neocortex sections from paraformaldehyde-fixed rat brains at embryonic day (E)16–18, and third instar larvae brains and optical discs from *Drosophila* flies expressing GFP-tagged centrosomin, were analysed by confocal scanning microscopy (Zeiss LSM 510 on an Axiovert 100M platform). For cells grown in culture, microscopy was performed with a Leica DMIRE2 microscope equipped with $\times 40$, $\times 63$ and $\times 100$ objectives, and images were captured using Leica Qfluoro software.

Live-cell imaging. Late/recycling endosomes were labelled with 10 mg ml^{-1} Lucifer Yellow (Sigma) for 10 min in a humid chamber at 37°C , rinsed, and allowed to internalize for periods ranging from 30 min to 2 h. Exocytic vesicles were labelled with $5 \mu\text{M}$ bodipy-ceramide (Molecular Probes) for 10 min. For live-cell analysis, neurons were grown on glass-bottomed Petri dishes (MaTek Corp.). *Drosophila* neuroblast division *in vitro* was performed as described²⁸, in cells from UASp-GFP–Cnn1 larvae.

CALI-mediated centrosome ablation. Neurons from *Drosophila* larvae expressing GFP-centrosomin¹² were used for centrosomal loss-of-function CALI (chromophore-assisted light inactivation) experiments^{15,16}. Neurons were seeded onto gridded coverslips for ulterior identification. A long-pass 570-nm filter was inserted in the illumination path to prevent excitation of GFP during cell identification. To irradiate the GFP-labelled centrosomes, we used different OD1 neutral density filters to obtain a power range from 55 W to 0.5 mW, as calculated in ref 15. Illumination was for 20 s, until the GFP emission was diminished. Changes in neuronal polarization and viability in both illuminated and non-illuminated cells were determined 24 h after excitation. CD8–GFP-expressing neurons (control neurons) were exposed to similar treatment and the same parameters were quantified.

Statistical analysis. Analysis of spatial correlation between the first lamellipodium and axon formation, or between the first lamellipodium, organelle clustering and axon formation was done as described in ref. 9. For analysis of the effects of nocodazole or brefeldin A, cells were grouped according to the presence or absence of neurites. The resulting populations were averaged and respective standard deviations pooled together. Student's *t*-tests were performed taking into account the two-tailed distribution and two-sample unequal variance of the pooled data.

Received 26 February; accepted 13 May 2005.

- da Silva, J. S. & Dotti, C. G. Breaking the neuronal sphere: regulation of the actin cytoskeleton in neuritogenesis. *Nature Rev. Neurosci.* **3**, 694–704 (2002).
- Horton, A. C. & Ehlers, M. D. Neuronal polarity and trafficking. *Neuron* **9**, 277–295 (2003).
- Cowan, C. R. & Hyman, A. A. Centrosomes direct cell polarity independently of microtubule assembly in *C. elegans* embryos. *Nature* **431**, 92–96 (2004).
- Singer, S. J. & Kupfer, A. The directed migration of eukaryotic cells. *Annu. Rev. Cell Biol.* **2**, 337–365 (1986).
- Dotti, C. G., Sullivan, C. A. & Banker, G. The establishment of polarity by hippocampal neurons in culture. *J. Neurosci.* **8**, 1454–1468 (1988).
- Bradke, F. & Dotti, G. C. The role of actin instability in axon formation. *Science* **283**, 1931–1934 (1999).
- Ueda, M., Graf, R., MacWilliams, H. K., Schliwa, M. & Euteneuer, U. Centrosome positioning and directionality of cell movements. *Proc. Natl Acad. Sci. USA* **94**, 9674–9678 (1997).
- Grill, S. W., Howard, J., Schaffer, E., Stelzer, E. H. & Hyman, A. A. The distribution of active force generators controls mitotic spindle position. *Science* **301**, 518–521 (2003).
- Dotti, C. G. & Banker, G. Intracellular organization of hippocampal neurons during the development of neuronal polarity. *J. Cell Sci.* **15** (suppl.), 75–84 (1991).
- Yonemura, E. M. *et al.* Rho-dependent transfer of citron-kinase to the cleavage furrow of dividing cells. *J. Cell Sci.* **114**, 3273–3284 (2001).
- Wu, C. F., Suzuki, N. & Poo, M. M. Dissociated neurons from normal and mutant *Drosophila* larval central nervous system in cell culture. *J. Neurosci.* **3**, 1888–1899 (1983).
- Megraw, T. L., Kilaru, S., Turner, F. R. & Kaufman, T. C. The centrosome is a dynamic structure that ejects PCM flares. *J. Cell Sci.* **115**, 4707–4718 (2002).
- Carney, G. E., Wade, A. A., Sapra, R., Goldstein, E. S. & Bender, M. *DHR3*, an ecdysone-inducible early-late gene encoding a *Drosophila* nuclear receptor, is required for embryogenesis. *Proc. Natl Acad. Sci. USA* **94**, 12024–12029 (1997).
- Houliston, E. & Maro, B. Post-translational modification of distinct microtubule subpopulations during cell polarization and differentiation in the mouse preimplantation embryo. *J. Cell Biol.* **108**, 543–551 (1989).
- Surrey, T. *et al.* Chromophore-assisted light inactivation and self-organization of microtubules and motors. *Proc. Natl Acad. Sci. USA* **14**, 4293–4298 (1998).
- Sydor, A. M., Su, A. L., Wang, F. S., Xu, A. & Jay, D. G. Talin and vinculin play distinct roles in filopodial motility in the neuronal growth cone. *J. Cell Biol.* **134**, 1197–1207 (1996).
- Wu, C. F., Sakai, K., Saito, M. & Hotta, Y. Giant *Drosophila* neurons differentiated from cytokinesis-arrested embryonic neuroblasts. *J. Neurobiol.* **21**, 499–507 (1990).
- Craig, A. M. & Banker, G. Neuronal polarity. *Annu. Rev. Neurosci.* **17**, 267–310 (1994).
- Yoshimura, T. *et al.* GSK-3 β regulates phosphorylation of CRMP-2 and neuronal polarity. *Cell* **14**, 137–149 (2005).
- Shi, S. H., Cheng, T., Jan, L. Y. & Jan, Y. N. APC and GSK-3 β are involved in mPar3 targeting to the nascent axon and establishment of neuronal polarity. *Curr. Biol.* **23**, 2025–2032 (2004).
- Schwamborn, J. C. & Puschel, A. W. The sequential activity of the GTPases Rap1B and Cdc42 determines neuronal polarity. *Nature Neurosci.* **7**, 923–929 (2004).
- Nishimura, T. *et al.* Role of the PAR-3–KIF3 complex in the establishment of neuronal polarity. *Nature Cell Biol.* **6**, 328–334 (2004).
- Shi, S. H., Jan, L. Y. & Jan, Y. N. Hippocampal neuronal polarity specified by spatially localized mPar3/mPar6 and PI 3-kinase activity. *Cell* **10**, 63–75 (2003).
- Goslin, K. & Banker, G. Experimental observations on the development of polarity by hippocampal neurons in culture. *J. Cell Biol.* **108**, 1507–1516 (1989).
- da Silva, J. S., Hasegawa, T., Miyagi, T., Dotti, C. G. & Abad-Rodríguez, J. Asymmetric membrane ganglioside sialidase activity specifies axonal fate. *Nature Neurosci.* **8**, 606–615 (2005).
- Piperno, G. & Fuller, M. T. Monoclonal antibodies specific for an acetylated form of α -tubulin recognize antigens in cilia and flagella from a variety of organisms. *J. Cell Biol.* **101**, 2085–2094 (1985).
- De Hoop, M., Meyn, L. & Dotti, C. G. In *Methods in Cell Biology: Laboratory Handbook* 2nd edn (ed. Celis, J. E.) 154–163 (Academic Press, San Diego, 1998).
- Feiguin, F., Llamazares, S. & Gonzalez, C. Methods in *Drosophila* cell cycle biology. *Curr. Top. Dev. Biol.* **36**, 279–291 (1998).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We would like to thank E. Cassin and B. Hellias for the hippocampal neurons, L. Ciapponi and the Bloomington Stock Center for fly strains, and C. Gonzalez for advice with the neuroblast division experiment. F.C.d.A. is supported by an EMBO long-term fellowship. J.S.D.S. was supported by an FCT/PRAXIS XXI scholarship (Portuguese Ministry of Science and Technology). F.F. was supported by an Alexander von Humboldt scholarship. Part of this work is supported by an EU Contract grant (APOPIIS) to C.G.D.

Author Contributions F.C.d.A. and G.P. were responsible for all *in vitro* experiments in mammalian and insect neurons, respectively. J.S.D.S. supervised the hippocampal neuron work. P.G.C. helped with the *in situ* work.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to F.F. (fabian.feiguin@unito.it) or C.G.D. (carlos.dotti@unito.it).

Licensing of natural killer cells by host major histocompatibility complex class I molecules

Sungjin Kim¹, Jennifer Poursine-Laurent¹, Steven M. Truscott², Lonnie Lybarger^{2†}, Yun-Jeong Song¹, Liping Yang¹, Anthony R. French^{1,3}, John B. Sunwoo^{1,4}, Suzanne Lemieux⁵, Ted H. Hansen² & Wayne M. Yokoyama^{1,2}

Self versus non-self discrimination is a central theme in biology from plants¹ to vertebrates, and is particularly relevant for lymphocytes that express receptors capable of recognizing self-tissues and foreign invaders. Comprising the third largest lymphocyte population, natural killer (NK) cells recognize and kill cellular targets and produce pro-inflammatory cytokines. These potentially self-destructive effector functions can be controlled by inhibitory receptors for the polymorphic major histocompatibility complex (MHC) class I molecules that are ubiquitously expressed on target cells^{2–4}. However, inhibitory receptors are not uniformly expressed on NK cells, and are germline-encoded by a set of polymorphic genes that segregate independently from MHC genes^{5,6}. Therefore, how NK-cell self-tolerance arises *in vivo* is poorly understood. Here we demonstrate that NK cells acquire functional competence through ‘licensing’ by self-MHC molecules. Licensing involves a positive role for MHC-specific inhibitory receptors and requires the cytoplasmic inhibitory motif originally identified in effector responses. This process results in two types of self-tolerant NK cells—licensed or unlicensed—and may provide new insights for exploiting NK cells in immunotherapy. This self-tolerance mechanism may be more broadly applicable within the vertebrate immune system because related germline-encoded inhibitory receptors are widely expressed on other immune cells.

NK cells from MHC-deficient mice and humans are defective in target killing^{7,8}. These defects could be due to effects on NK cell subpopulations, or altered expression or function of activation or inhibitory receptors^{9,10}. These issues are difficult to assess with cellular targets because the repertoire of receptors on NK cells that are involved in target recognition is incompletely defined. Here we describe the use of a target cell-free system (see Methods) to analyse the activation of individual resting NK cells by stimulation through the NK1.1 molecule, also known as NK-cell receptor protein 1c (Nkrp1c) or killer-cell lectin-like receptor subfamily B member 1c (Klrb1c). NK1.1 is an activation receptor expressed by all mature and developing CD3⁺ NK cells in C57BL/6J (B6) mice¹¹. NK1.1 cross-linking led to the production of interferon (IFN)- γ by large numbers of NK cells from wild-type B6 mice, but not from MHC class Ia- and Ib-deficient mice that lack β_2 -microglobulin ($\beta_2m^{-/-}$), or from mice that lack MHC class Ia alleles only ($H2-K^{b-/-};D^{b-/-}$) (Fig. 1a). Even with high anti-NK1.1 antibody concentrations, many fewer $\beta_2m^{-/-}$ or $K^{b-/-};D^{b-/-}$ NK cells produced IFN- γ (Supplementary Fig. 1). Similar effects were observed with stimulation through other activation receptors and by tumour targets, although no differences were noted when the NK cells were stimulated with phorbol 12-myristate 13-acetate (PMA) and ionomycin

(Supplementary Figs 2 and 3). However, the differential activation of NK cells from MHC-sufficient and MHC-deficient hosts was less pronounced when NK cells were pre-stimulated with interleukin-2 (IL-2) or polyinosinic-polycytidylic acid (poly-I:C) (Supplementary Fig. 4). Finally, co-incubation of wild-type and $\beta_2m^{-/-}$ splenocytes during NK1.1 crosslinking led to IFN- γ production selectively from wild-type NK cells (data not shown), indicating a cell-autonomous effect. Thus, the intrinsic defects of NK cells from MHC class Ia-deficient hosts suggest that NK cell functional maturation requires specific interaction with host MHC class Ia molecules. We term this process MHC class I-dependent ‘licensing’ to distinguish it from MHC class I-dependent ‘education’, which implies different events occurring during T-cell development.

The Ly49 C-type lectin family of receptors (now also known as killer cell lectin-like receptors subfamily A or Klra) that are expressed on NK cells may be involved in licensing because they recognize MHC class Ia molecules in effector responses^{2,12,13}. However, licensed NK cells were found in mice deficient in DAP12 (DNAX activating protein of 12 kDa; also known as killer-cell activating receptor-associated protein or KARAP) or other signalling-chain molecules (DAP10, Fc ϵ RI γ or CD3 ζ) (Supplementary Fig. 2), indicating that Ly49 (ref. 14) and other activation receptors that are associated with these adaptor molecules are not critical for licensing. As for Ly49 inhibitory receptors, we analysed NK cells expressing Ly49A and Ly49C because of their well-defined MHC class I specificities and the availability of monoclonal antibodies¹². Ly49C⁺ NK cells from wild-type B6 ($H2^b$) mice readily produced IFN- γ in response to NK1.1 stimulation, but not when they were derived from $\beta_2m^{-/-}$ or $K^{b-/-};D^{b-/-}$ mice (Fig. 1a). This effect was not due to differences in NK1.1 expression between Ly49C⁺ and Ly49C[−] subsets, or between wild-type and MHC-deficient NK cells (data not shown). In contrast, Ly49A⁺ NK cells from wild-type B6 mice showed poor IFN- γ production. We therefore postulated that licensing was related to self-H2 ligands for specific Ly49 receptors because Ly49A does not have an $H2^b$ ligand, whereas Ly49C can recognize many MHC class I ligands including $H2-K^b$ (refs 2, 12). Indeed, stimulated Ly49A⁺ NK cells strongly produced IFN- γ only when they were derived from MHC-congenic B10 mouse strains that express the Ly49A ligand $H2-D^d$ (Fig. 1b). Production of tumour necrosis factor (TNF)- α followed a similar pattern (Fig. 1b). Likewise, resting Ly49A⁺ NK cells from B10.D2 mice deficient in recombination-activating gene 2 (B10.D2 $Rag2^{-/-}$) killed YAC-1 and CHO target cells better than Ly49A[−] NK cells from the same animals (Fig. 1c). In contrast, Ly49A⁺ NK cells from B10 $Rag2^{-/-}$ mice killed with similar efficiency to Ly49A[−] NK cells from the same animals. Taken together, these data indicate that NK-cell competence for cytokine production

¹Howard Hughes Medical Institute, Rheumatology Division, Departments of Medicine, ²Pathology and Immunology, ³Pediatrics and ⁴Otolaryngology, Washington University School of Medicine, St. Louis, Missouri 63110, USA. ⁵Institut national de la recherche scientifique, INRS-Institut Armand-Frappier, Université du Québec, Laval, Québec H7V 1B7, Canada. [†]Present address: Department of Cell Biology and Anatomy, College of Medicine, University of Arizona, Tucson, Arizona 85724, USA.

and cytolysis is dependent on host MHC class I molecules. This positive effect is associated with the expression of self-MHC class I-specific receptors on NK cells that were originally identified as inhibitory receptors in effector responses².

We studied NK cells derived from *H2-D^d* transgenic *K^b^{-/-};D^b^{-/-}* mice that express only one MHC class Ia molecule. Ly49A⁺ NK cells from these mice were readily triggered by anti-NK1.1, whereas cells from *K^b^{-/-};D^b^{-/-}* mice were not (Fig. 2a). Similar effects were observed with NK cells derived from B6 (*H2^b*) mice that were transgenic for *H2-D^d* (data not shown). The effect of transgenic *H2-D^d* molecules in *K^b^{-/-};D^b^{-/-}* mice was also manifested in the Ly49A⁻ population (Fig. 2a), and was perhaps due to the presence of other Ly49 receptors with *H2-D^d* specificity¹², recognition by the NKG2/CD94 heterodimer of the MHC class I molecule Qa-1 (which

can present the *H2-D^d* leader peptide to NK cells)¹⁵, or an unknown receptor on NK cells for MHC class I molecules. Nevertheless, the data indicate that NK cells expressing a cognate inhibitory receptor for a self-MHC class I molecule possess enhanced functional activities.

To more precisely investigate the functional relationship between Ly49 inhibitory receptors and host MHC class I molecules, we used a single-chain trimer (SCT) MHC class I molecule consisting of peptide, β_2m and *H2-K^b* heavy chain connected by Gly-Ser linkers¹⁶. SCT-*K^b* tetramers bound a subset of β_2m ^{-/-} NK cells that co-stained with the monoclonal antibody 4L03311 (Fig. 2b), which is monospecific for Ly49C (ref. 17). This result was similar to the staining of NK cells from MHC class I-disparate mice with conventional MHC class I tetramers¹⁸. In contrast, staining with other Ly49 monoclonal antibodies did not correlate with tetramer binding. Furthermore, SCT-*K^b* tetramer binding was specifically blocked by pre-incubation with anti-Ly49C (Fig. 2c). Similar data were obtained with conventional (wild-type) *K^b* tetramers (data not shown) that bind Ly49C transfectants¹².

To generate an SCT-*K^b* transgenic mouse that does not express any other β_2m -associated MHC class Ia or Ib molecules, we produced transgenic mice expressing SCT-*K^b* ubiquitously (data not shown), and backcrossed the founder mouse to triple-knockout β_2m ^{-/-}; *K^b^{-/-}*; *D^b^{-/-}* animals. The *cis* interactions between an NK-cell receptor and its cognate MHC ligand were recently reported to block MHC class I tetramer binding, providing evidence for *in vivo* self-MHC ligand-receptor engagement¹⁸. In SCT-*K^b*-transgenic and wild-type B6 mice, binding of both SCT-*K^b* and conventional (wild-type) *H2-K^b* tetramers was similarly absent, indicating that this biological property was intact with the transgenic SCT-*K^b* molecules (Fig. 2d). Finally, anti-Ly49C reactivity was lower in SCT-*K^b*-transgenic β_2m ^{-/-}; *K^b^{-/-}*; *D^b^{-/-}* mice compared with non-transgenic β_2m ^{-/-}; *K^b^{-/-}*; *D^b^{-/-}* mice (Fig. 2e), and is consistent with decreased antibody reactivity to NK-cell receptors in the presence of self-MHC class I ligands^{9,19}. Thus, the SCT-*K^b* molecules react specifically with Ly49C on primary NK cells and exhibit biological properties similar to wild-type *H2-K^b* molecules.

Ly49C⁺ NK cells from the SCT-*K^b*-transgenic β_2m ^{-/-}; *K^b^{-/-}*; *D^b^{-/-}* mice gained the ability to produce IFN- γ , compared with Ly49C⁺ NK cells from non-transgenic triple-knockout mice (Fig. 2e), whereas Ly49C⁻ NK cells, even with higher anti-NK1.1 stimulation, were comparable between transgenic and non-transgenic triple-knockout mice. Moreover, there was no specific enhancement of Ly49A⁺ NK-cell activity in SCT-*K^b*-transgenic triple-knockout mice (data not shown). Thus, the expression of a single MHC class I molecule selectively licenses functional activity of NK cells with a cognate inhibitory receptor.

To further study the action of NK-cell receptors in licensing, we transduced haematopoietic stem cells with bicistronic retroviral vectors containing green fluorescent protein (GFP) to distinguish expression of transduced Ly49A from endogenous Ly49A. As expected, there was correlated expression of Ly49A and GFP (Supplementary Fig. 5). In B10.D2 mice, GFP⁺ NK cells were more potent at producing IFN- γ compared with GFP⁻ NK cells (Fig. 3a). The relative frequencies of IFN- γ -producing cells were about twofold higher in GFP⁺ cells compared with GFP⁻ cells (Fig. 3c). In contrast, this effect was not seen in B10 hosts (Fig. 3a). Thus, transduced Ly49A expression conferred a positive effect on NK cell licensing that exceeded the existing NK-cell function, but only in hosts with a cognate MHC class I ligand for Ly49A.

Licensing by Ly49A could be due to a signal from Ly49A itself. Alternatively, Ly49A could activate stromal cells through MHC crosslinking, or facilitate cell-cell contact between developing NK cells and stromal elements. To address these possibilities, we expressed a Ly49A molecule that lacked the cytoplasmic domain (designated Ly49A^{cytoΔ}) in B10.D2 mice. Although this molecule was

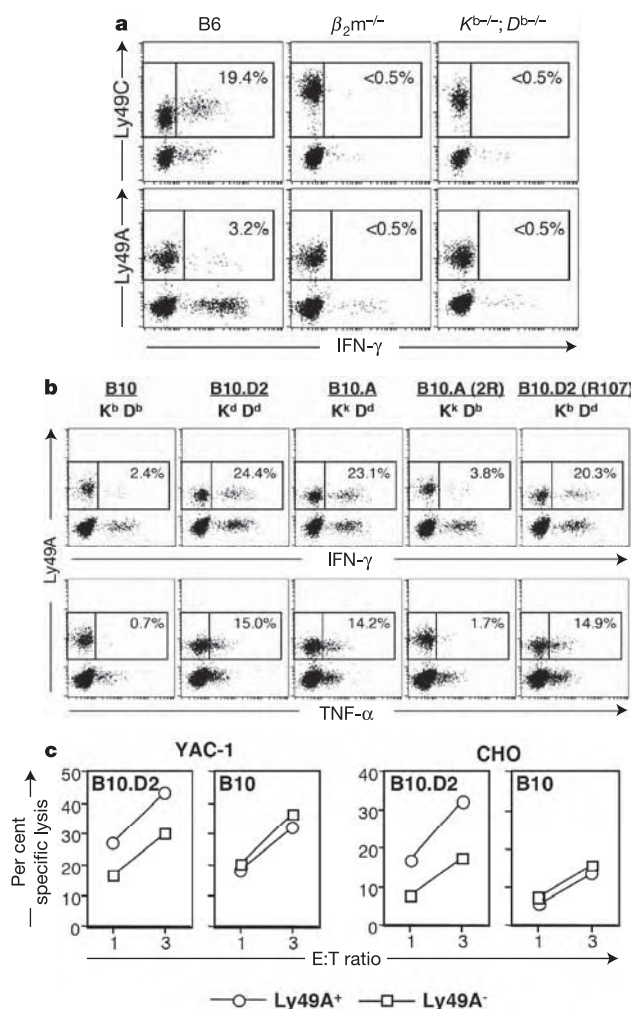


Figure 1 | Host MHC class Ia molecules alter IFN- γ production and cytotoxic activity of NK subsets. **a**, Splenocytes from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Gated CD3⁺ TCR- β ⁺ DX5⁺ cells are shown, and the numbers represent the percentages of IFN- γ ⁺ cells among the Ly49A⁺ or Ly49C⁺ cell populations. **b**, Splenocytes from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ and TNF- α . The H2 MHC haplotypes are indicated under each strain name. Gated CD3⁺ TCR- β ⁺ DX5⁺ cells are shown, and the numbers represent the percentages of cytokine-producing cells among the Ly49A⁺ cell population. **c**, Ly49A⁺ NK1.1⁺ and Ly49A⁻ NK1.1⁺ cells were sorted from the spleen cells of *Rag2*-deficient B10.D2 and *Rag2*-deficient B10 mice, and used in standard 4-h killing assays against the target cells indicated. Effector:target (E:T) ratio is indicated on the x axis.

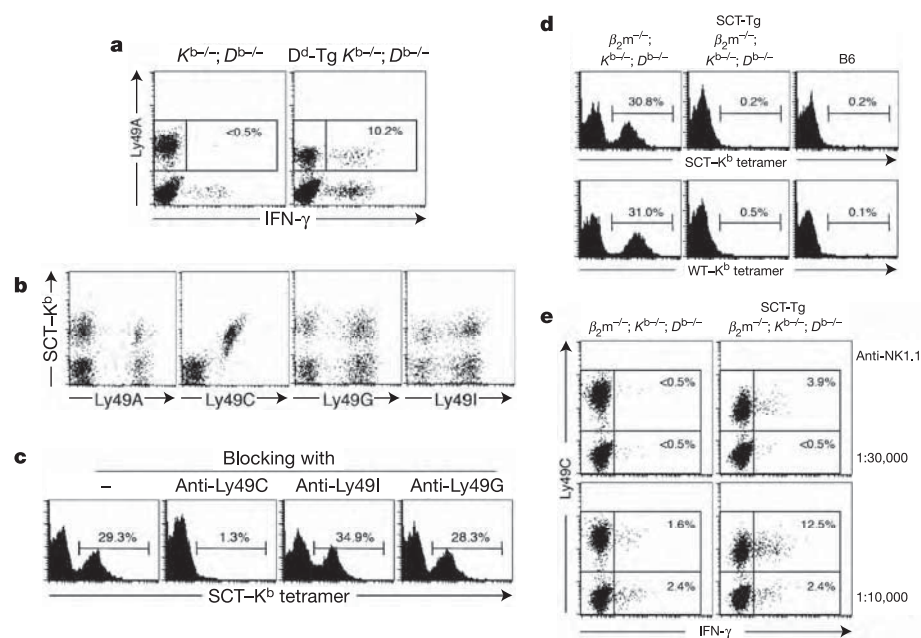


Figure 2 | Licensing of NK cells expressing inhibitory Ly49 receptors specific for self-MHC class I molecules. **a**, Splenocytes from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Gated CD3 $^{-}$ TCR- β $^{-}$ DX5 $^{+}$ cells are shown, and the numbers represent the percentages of IFN- γ $^{+}$ cells among the Ly49A $^{+}$ cell population. **b**, $\beta_2m^{-/-}$ splenocytes were stained with the indicated monoclonal antibodies and SCT-K b tetramer together. Gated CD3 $^{-}$ CD19 $^{-}$ NK1.1 $^{+}$ cells are shown. **c**, $\beta_2m^{-/-}$ splenocytes were first incubated with the indicated monoclonal antibodies and subsequently stained with SCT-K b tetramer. Gated CD3 $^{-}$ CD19 $^{-}$ NK1.1 $^{+}$ cells are shown. **d**, Splenocytes from the indicated mouse strains were stained with SCT-K b or WT-K b tetramer. Gated CD3 $^{-}$ CD19 $^{-}$ NK1.1 $^{+}$ cells are shown. **e**, Splenocytes from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody (1:30,000 and 1:10,000 dilution) and analysed for intracellular IFN- γ . Gated CD3 $^{-}$ TCR- β $^{-}$ DX5 $^{+}$ cells are shown, and the numbers represent the percentages of IFN- γ $^{+}$ cells among the Ly49C $^{+}$ or Ly49C $^{-}$ cell populations. Tg, transgenic.

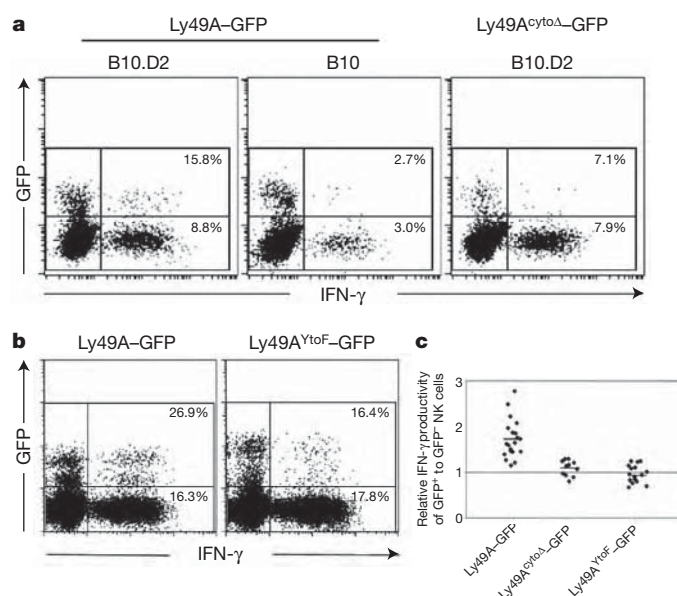


Figure 3 | Gene transfer of intact Ly49A, but not cytoplasmic domain-deleted (Ly49A $^{cyto\Delta}$) or ITIM-mutated (Ly49A YtoF) Ly49A, licenses NK cells in the presence of its ligand. **a**, Mice were reconstituted with syngeneic BM cells after retroviral transduction of the indicated constructs. Splenic cells from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Gated CD3 $^{-}$ TCR- β $^{-}$ DX5 $^{+}$ cells are shown, and the numbers represent the percentages of IFN- γ $^{+}$ cells among the GFP $^{+}$ or GFP $^{-}$ cell populations. **b**, Splenic cells from B10.D2 mice reconstituted with syngeneic BM cells after retroviral transduction of the indicated constructs were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Gated CD3 $^{-}$ TCR- β $^{-}$ CD19 $^{-}$ NK1.1 $^{+}$ cells are shown, and the numbers represent the percentages of IFN- γ $^{+}$ cells among GFP $^{+}$ or GFP $^{-}$ cells. **c**, Relative IFN- γ productivity is depicted as the ratio of the percentage of IFN- γ $^{+}$ cells among the GFP $^{+}$ NK cells to the percentage of IFN- γ $^{+}$ cells among the GFP $^{-}$ NK cells. Each symbol represents an individual mouse and the bars represent the mean ratios of the indicated groups; wild-type Ly49A (mean $\pm$ s.d., 1.73 ± 0.42 , $n = 20$), Ly49A $^{cyto\Delta}$ (1.09 ± 0.17 , $n = 11$) and Ly49A YtoF (0.96 ± 0.19 , $n = 17$). The differences between wild-type Ly49A-GFP and either Ly49A $^{cyto\Delta}$ -GFP or Ly49A YtoF -GFP were statistically significant ($P < 0.0001$).

expressed at comparable levels (Supplementary Fig. 5), we did not observe enhanced NK-cell activity (Fig. 3a, c). The only known signalling motif in the Ly49A cytoplasmic domain is the immunoreceptor tyrosine-based inhibitory motif, or ITIM (ref. 20). A mutant Ly49A (designated Ly49A YtoF) with a Tyr-to-Phe change in the ITIM, known to abrogate inhibitory signalling in effector responses²⁰, was expressed by retroviral transduction at levels comparable to intact Ly49A (Supplementary Fig. 5). The ITIM mutant failed to confer licensing in B10.D2 mice (Fig. 3b, c). Taken together, these results show that the ITIM in the cytoplasmic domain of the inhibitory receptor is critical for the licensing effect, and that the receptor itself is responsible for the signals that direct NK-cell licensing.

The ITIM of Ly49A recruits the tyrosine phosphatase SHP1 (src homology region 2 domain-containing phosphatase 1; also known as protein tyrosine phosphatase non-receptor type 6 or PTPN6), which is required for inhibitory signalling in effector responses²⁰. To examine its role in licensing, we analysed NK-cell function in SHP1-defective motheaten-viable (*me-v*) mice (*Ptpn6* $^{me-v}$). NK1.1 crosslinking led to IFN- γ production from *me-v*-NK cells at levels that were intermediate between β_2m -deficient and wild-type NK cells (Fig. 4a). To address whether the results were influenced by the pleiotropic abnormalities found in *me-v* mice, we studied mixed-bone-marrow (BM) chimaeras, which use different alleles of Ly5 (also known as protein tyrosine phosphatase receptor type C or PTPRC) to distinguish BM origin. BM derived from *me-v* mice (carrying the Ly5.2 allele) and wild-type mice (carrying the Ly5.2 or the Ly5.1 allele) was transplanted into irradiated wild-type recipient mice (carrying Ly5.1). In the mixed-chimaeric mice, *me-v*-derived and wild-type-derived NK cells displayed similar levels and patterns of IFN- γ production (Fig. 4b), indicating that the originally observed reduced IFN- γ production by *me-v* NK cells (Fig. 4a) was due to indirect effects. ITIMs can potentially recruit other molecules, but preliminary data do not implicate known interacting molecules in licensing (Supplementary Note). Nevertheless, these data suggest that although licensing is mediated through an ITIM-dependent mechanism, it is distinct from ITIM-mediated, SHP1-dependent inhibition during effector responses.

Here we show that NK cells, like other lymphocytes, undergo a process that results in self-tolerance. Consistent with a positive effect during development, BM-derived NK cells show evidence of enhanced proliferation if they express an inhibitory receptor for a

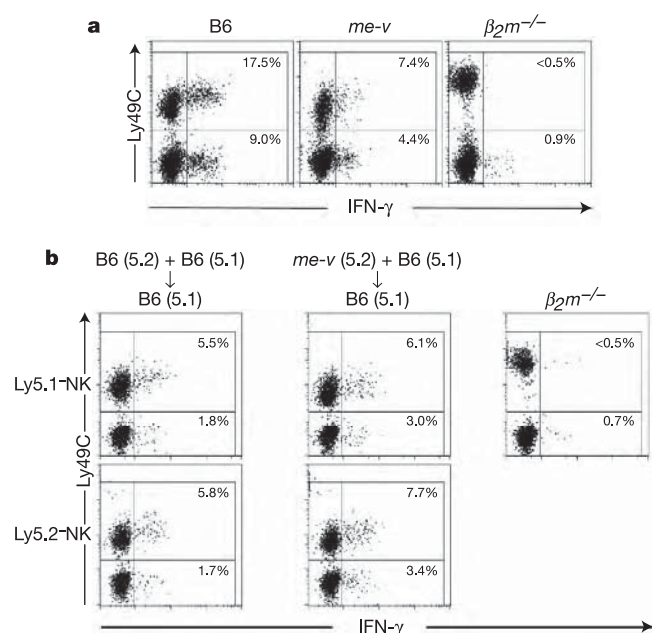


Figure 4 | Licensing is preserved in SHP1-deficient NK cells. **a**, Splenocytes from the indicated mouse strains were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Gated CD3⁺ TCR- β ⁺ CD19⁺ NK1.1⁺ cells are shown, and the numbers represent the percentages of IFN- γ ⁺ cells among the Ly49C⁺ or Ly49C⁻ cells. **b**, Splenocytes from mixed BM chimaeric mice or from $\beta_2m^{-/-}$ control mice were incubated with immobilized anti-NK1.1 antibody and analysed for intracellular IFN- γ . Mixed BM chimaeric mice were produced by injection of Ly5.2⁺ B6 plus Ly5.1⁺ B6 or Ly5.2⁺ *me-v* plus Ly5.1⁺ B6 BM into irradiated Ly5.1⁺ B6 mice and analysed after 9 weeks. Donor-derived wild-type or *me-v* NK cells were distinguished by the presence of the different Ly5 alleles. Gated Ly5.1⁺ CD3⁺ TCR- β ⁺ CD19⁺ NK1.1⁺ (upper panel) and Ly5.2⁺ CD3⁺ TCR- β ⁺ CD19⁺ NK1.1⁺ (lower panel) cells are shown, and the numbers represent the percentages of cytokine-producing cells among the Ly49C⁺ or Ly49C⁻ cell populations. Data shown are representative of five mice per group analysed in two separate experiments.

self-MHC class I molecule (Supplementary Fig. 6). In some respects, licensing resembles the process of positive selection that occurs in the thymus for T cells with appropriate self-reactive T-cell antigen receptors²¹. However, NK cells use a distinctly different process that is ITIM-dependent, and there is no evidence, as yet, for negative

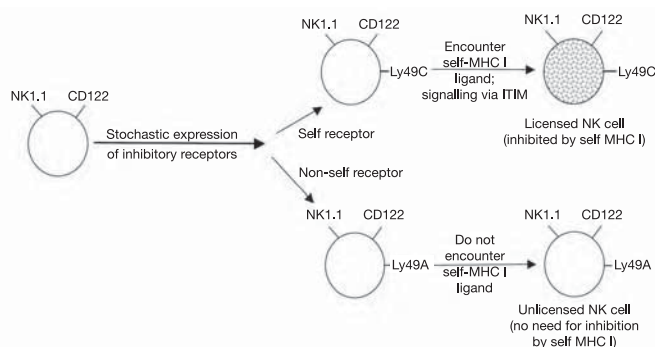


Figure 5 | Model for NK cell licensing in the H2-K^b+ D^d- mouse. For clarity, this diagram depicts the situation for developing NK cells bearing either Ly49C alone (top) or Ly49A alone (bottom). Licensing results in two types of self-tolerant NK cells. The Ly49C⁺ NK cell becomes licensed and is inhibited by H2-K^b (self), whereas cells without self-specific receptors are not licensed even though they can express receptors for other MHC alleles; that is, the Ly49A⁺ NK cell is not licensed.

selection or clonal deletion of NK cells. In contrast to T cells, which use rearranged receptors that activate effector functions in adaptive immunity, NK cells use non-rearranged receptors that are, paradoxically, involved in effector inhibition to find self antigens for functional maturation in innate immunity (Fig. 5). Therefore, licensing results in two types of self-tolerant NK cells—licensed NK cells with self-MHC class I inhibitory receptors, and unlicensed NK cells.

Licensing pairs an inhibitory receptor with its cognate self-MHC class I ligand for functional development, thus providing a mechanism for how independently segregating genes for non-rearranged receptors and ligands give rise to tolerance of self antigens. NK cells that do not express self-MHC class I specific receptors do not become licensed, and have no need to be inhibited by MHC class I because they are not functionally competent. Therefore, MHC class I-dependent effector inhibition is relevant only for competent NK cells; they are inhibited by the same self-MHC-specific receptors that originally conferred licensing. Finally, germline-encoded inhibitory receptors with ITIMs, related to the NK cell receptors, are broadly expressed on many leukocytes²², suggesting that they might also be involved in self-tolerance mechanisms.

It should be noted that each individual NK cell is potentially licensed separately; in hosts with heterozygous MHC alleles, there should be individual NK cells that are licensed and thus inhibited by different MHC molecules. In this respect, licensing should be relevant to understanding the function of NK cells in other contexts, such as BM transplantation, and the vexing problem of 'hybrid resistance' in which host NK cells in F₁-hybrid mice reject BM from inbred parental strains²³. Only normal tissues expressing the full complement of self-MHC class I molecules should be resistant to all licensed NK cells in the F₁ hybrid, whereas tissues bearing only one haplotype would be resistant to some, but not all, NK cells. On the other hand, tissues expressing non-self MHC molecules that are promiscuously recognized by the inhibitory receptors on licensed NK cells should be protected—an important issue in interpreting results from BM transplant studies in mice and humans²⁴. In related studies, donor-derived NK cells have an anti-tumour effect in BM transplantation therapy for leukaemia²⁵. Although recipient HLA class I alleles are clearly important, the absence of the appropriate recipient HLA ligands for the donor killer cell immunoglobulin-like inhibitory receptors (KIRs; functionally similar to murine Ly49 receptors) does not always impart an effective outcome²⁶. Licensing indicates that donor HLA class I alleles will influence the functional capacity of transplanted NK cells.

Notably, the licensing effect was less prominent in pre-activated NK cells, suggesting that licensing requirements could be bypassed under certain circumstances. Perhaps unlicensed NK cells become activated in response to pro-inflammatory cytokines, which are known to activate most, if not all, NK cells during infections²⁷. Such effects may aid anti-pathogen defences by potentially breaking NK-cell tolerance and allowing the participation of larger numbers of NK cells at local sites of inflammation, suggesting a potentially important function of unlicensed NK cells.

METHODS

Mice. C57BL/6J mice (wild-type B6 mice), B6 mice carrying the Ly5.1 antigen, and C57BL/10J (B10) congenic mice (including B10.D2, B10.A, B10.A(2R) and B10.D2(R107)) were purchased from The Jackson Laboratory. In addition, $\beta_2m^{-/-}$, *FcεRIγ*^{-/-}, *CD3ζ*^{-/-} and motheaten-viable *Ptpn6*^{me-v} mice, bred onto the B6 background, were also purchased from The Jackson Laboratory. B10 *Rag2*^{-/-} and B10.D2 *Rag2*^{-/-} mice were purchased from Taconic Farms. D^d-transgenic B6 mice were provided by D. Margulies, and crossed to the K^b-/-;D^b-/- double-knockout mice provided by F. Lemonnier. *DAP12*^{-/-} and *DAP10*^{-/-} mice were provided by T. Takai and M. Colonna, respectively, and were completely backcrossed to the B6 background using a marker-assisted breeding strategy. All mice were used in accordance with institutional guidelines for animal experimentation.

SCT transgenic mice. A transgene construct for the expression of the ovalbumin-peptide (OVA)-K^b SCT (see ref. 28 for a detailed description of the SCT) was generated from an H2-K^b genomic clone. The first and second exons of the H2-K^b gene were replaced with a hybrid exon encoding the leader sequence of murine β_2m , followed by the OVA-derived SIINFEKL peptide, a 15-mer Gly-Ser linker, the mature coding sequence of murine β_2m , a 20-mer Gly-Ser linker and the $\alpha 1$ domain of H2-K^b. SCT-K^b is recognized by antigen-specific, H2-K^b-restricted cytotoxic T lymphocytes, indicating that its three-dimensional topology is preserved¹⁶. Expression of the OVA-K^b transgene was driven by a 2-kilobase fragment encompassing the H2-L^d promoter, and microinjection of the purified expression cassette into B6 embryos was performed by the Mouse Genetics Core at Washington University. Transgene-positive founders were crossed to B6 mice with targeted disruptions of the β_2m and K^b;D^b genes²⁹.

Cytokine assays. Spleen cells were collected from mice that did not receive any prior treatment, unless otherwise indicated. For the stimulation of NK cells, spleen cells (10^7 cells ml⁻¹) were incubated with an equal volume of tumour targets (10^6 cells ml⁻¹) for 30–60 min, and then further incubated in the presence of brefeldin A for an additional 5–8 h. Alternatively, spleen cells (10^7 cells ml⁻¹) were incubated in 6-well plates coated with anti-NK1.1 monoclonal antibody ascites at a 1:30,000 dilution (chosen for most experiments because its effects were comparable to tumour stimulation; see Supplementary Fig. 2), unless otherwise indicated. Similar observations were made with purified monoclonal antibody (data not shown). Intracellular IFN- γ was detected with either phycoerythrin (PE)-conjugated or fluorescein isothiocyanate (FITC)-conjugated anti-IFN- γ monoclonal antibody (XMG1.2) as described previously¹¹. Alternatively, intracellular TNF- α was detected with PE-conjugated anti-TNF- α antibody on cells stimulated with anti-NK1.1 at 1:10,000 dilution. To gate on NK cells, cells were stained for either the universal NK cell marker DX5 or NK1.1, which yielded essentially identical results. To exclude T cells from analysis, both CD3⁺ and T-cell antigen receptor (TCR)- β ⁺ cells were gated out in all experiments, and CD19⁺ cells were gated out in some experiments.

Cytotoxicity assay. Ly49A⁺ and Ly49A⁻ NK cells were sorted from the spleen cells of unmanipulated Rag2-deficient B10 and Rag2-deficient B10.D2 mice to prevent inadvertent contamination with NKT cells. Sorted cells (with a comparable purity of >90%) were tested directly in standard ⁵¹Cr-release assays using 96-well V-bottom plates.

Retroviral gene transduction. Retroviral vectors pMX-Ly49A-IRES-EGFP, pMX-Ly49A^{cytoA}-IRES-EGFP and pMX-Ly49A^{cytoF}-IRES-EGFP were constructed by inserting the Ly49A complementary DNA, its cytoplasmic tail-deleted mutant cDNA, or the Ly49A^{cytoF} mutant cDNA into the pMX-IRES-GFP vector (a gift from T. Kitamura), respectively. The mutants were generated by polymerase chain reaction mutagenesis and verified by sequence analysis. The cytoplasmic deletion mutant had an amino-terminal deletion (amino acids 2 to 32), whereas the Ly49A^{cytoF} mutant had a codon mutation (TAT to TTC at codon 8). The vectors were transfected into the Plat-E packaging cell line with the FuGENE 6 transfection system (Roche) to produce recombinant virus. The viral supernatants were used to infect BM haematopoietic precursors as described previously³⁰. Infected BM cells were injected into the tail vein of 9.5-Gy-irradiated B10 or B10.D2 mice. The spleen cells were analysed at least 6 weeks after transplantation.

Mixed bone marrow chimaeras. BM cells were prepared from femora and tibiae of wild-type B6 (Ly5.2), *me-v* (Ly5.2) and wild-type B6 (Ly5.1) mice. A total of 2×10^6 mixed BM cells (Ly5.2⁺B6 plus Ly5.1⁺B6; or Ly5.2⁺ *me-v* plus Ly5.1⁺B6), mixed at a 4:1 (Ly5.2:Ly5.1) ratio, were injected via the tail vein of 9.5-Gy-irradiated wild-type B6 (Ly5.1) mice. The spleen cells were analysed at least 9 weeks after transplantation.

Received 27 March; accepted 19 May 2005.

- Murase, K. *et al.* A membrane-anchored protein kinase involved in Brassica self-incompatibility signalling. *Science* **303**, 1516–1519 (2004).
- Karlhofer, F. M., Ribaldo, R. K. & Yokoyama, W. M. MHC class I alloantigen specificity of Ly-49⁺ IL-2-activated natural killer cells. *Nature* **358**, 66–70 (1992).
- Lanier, L. L. NK cell receptors. *Annu. Rev. Immunol.* **16**, 359–393 (1998).
- Moretta, L., Biassoni, R., Bottino, C., Mingari, M. C. & Moretta, A. Human NK-cell receptors. *Immunol. Today* **21**, 420–422 (2000).
- Trowsdale, J. Genetic and functional relationships between MHC and NK receptor genes. *Immunity* **15**, 363–374 (2001).
- Yokoyama, W. M. & Plougastel, B. F. Immune functions encoded by the natural killer gene complex. *Nature Rev. Immunol.* **3**, 304–316 (2003).

- Bix, M. *et al.* Rejection of class I MHC-deficient haematopoietic cells by irradiated MHC-matched mice. *Nature* **349**, 329–331 (1991).
- Furukawa, H. *et al.* Tolerance of NK and LAK activity for HLA class I-deficient targets in a TAP1-deficient patient (bare lymphocyte syndrome type I). *Hum. Immunol.* **60**, 32–40 (1999).
- Raulet, D. H., Vance, R. E. & McMahon, C. W. Regulation of the natural killer cell receptor repertoire. *Annu. Rev. Immunol.* **19**, 291–330 (2001).
- Sentman, C. L., Olsson, M. Y. & Karre, K. Missing self recognition by natural killer cells in MHC class I transgenic mice. A 'receptor calibration' model for how effector cells adapt to self. *Semin. Immunol.* **7**, 109–119 (1995).
- Kim, S. *et al.* *In vivo* developmental stages in murine natural killer cell maturation. *Nature Immunol.* **3**, 523–528 (2002).
- Hanke, T. *et al.* Direct assessment of MHC class I binding by seven Ly49 inhibitory NK cell receptors. *Immunity* **11**, 67–77 (1999).
- Nakamura, M. C. *et al.* Mouse Ly-49D recognizes H2-D^d and activates natural killer cell cytotoxicity. *J. Exp. Med.* **189**, 493–500 (1999).
- Bakker, A. B. *et al.* DAP12-deficient mice fail to develop autoimmunity due to impaired antigen priming. *Immunity* **13**, 345–353 (2000).
- Vance, R. E., Kraft, J. R., Altman, J. D., Jensen, P. E. & Raulet, D. H. Mouse CD94/NKG2A is a natural killer cell receptor for the nonclassical major histocompatibility complex (MHC) class I molecule Qa-1(b). *J. Exp. Med.* **188**, 1841–1848 (1998).
- Lybarger, L. *et al.* Enhanced immune presentation of a single-chain major histocompatibility complex class I molecule engineered to optimize linkage of a C-terminally extended peptide. *J. Biol. Chem.* **278**, 27105–27111 (2003).
- Brennan, J., Lemieux, S., Freeman, J. D., Mager, D. L. & Takei, F. Heterogeneity among Ly-49C natural killer (NK) cells—characterization of highly related receptors with differing functions and expression patterns. *J. Exp. Med.* **184**, 2085–2090 (1996).
- Doucey, M. A. *et al.* Cis association of Ly49A with MHC class I restricts natural killer cell inhibition. *Nature Immunol.* **5**, 328–336 (2004).
- Andersson, M. *et al.* MHC class I mosaic mice reveal insights into control of Ly49C inhibitory receptor expression in NK cells. *J. Immunol.* **161**, 6475–6479 (1998).
- Nakamura, M. C. *et al.* Mouse Ly-49A interrupts early signalling events in natural killer cell cytotoxicity and functionally associates with the Shp-1 tyrosine phosphatase. *J. Exp. Med.* **185**, 673–684 (1997).
- Starr, T. K., Jameson, S. C. & Hogquist, K. A. Positive and negative selection of T cells. *Annu. Rev. Immunol.* **21**, 139–176 (2003).
- Long, E. O. Regulation of immune responses through inhibitory receptors. *Annu. Rev. Immunol.* **17**, 875–904 (1999).
- Bennett, M. *et al.* Hybrid resistance: 'negative' and 'positive' signalling of murine natural killer cells. *Semin. Immunol.* **7**, 121–127 (1995).
- Barao, I. & Murphy, W. J. The immunobiology of natural killer cells and bone marrow allograft rejection. *Biol. Blood Marrow Transplant.* **9**, 727–741 (2003).
- Ruggeri, L. *et al.* Effectiveness of donor natural killer cell alloreactivity in mismatched hematopoietic transplants. *Science* **295**, 2097–2100 (2002).
- Parham, P. & McQueen, K. L. Alloreactive killer cells: hindrance and help for haematopoietic transplants. *Nature Rev. Immunol.* **3**, 108–122 (2003).
- Biron, C. A., Nguyen, K. B., Pien, G. C., Cousens, L. P. & Salazar-Mather, T. P. Natural killer cells in antiviral defense: function and regulation by innate cytokines. *Annu. Rev. Immunol.* **17**, 189–220 (1999).
- Yu, Y. Y., Netuschil, N., Lybarger, L., Connolly, J. M. & Hansen, T. H. Cutting edge: single-chain trimers of MHC class I molecules form stable structures that potentially stimulate antigen-specific T cells and B cells. *J. Immunol.* **168**, 3145–3149 (2002).
- Lybarger, L., Wang, X., Harris, M. R., Virgin, H. W. & Hansen, T. H. Virus subversion of the MHC class I peptide-loading complex. *Immunity* **18**, 121–130 (2003).
- Furukawa, H., Iizuka, K., Poursine-Laurent, J., Shastri, N. & Yokoyama, W. M. A ligand for the murine NK activation receptor Ly-49D: activation of tolerized NK cells from β_2 -microglobulin-deficient mice. *J. Immunol.* **169**, 126–136 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements Work in the Yokoyama laboratory is supported by the Howard Hughes Medical Institute, the Barnes-Jewish Hospital Foundation and grants from the National Institutes of Health. Transgenic production and genotyping were supported by the Rheumatic Diseases Core Center grant. This study was also supported by an NIH grant to the Hansen laboratory. The authors thank M. Miley and D. Fremont for initial production of the SCT-K^b tetramers, E. Holroyd, J. Mohan, D. Higuchi and R. Rodrigues for technical assistance, and P. Allen, M. Colonna, J. Loh and E. Unanue for critical comments on the manuscript.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to W.M.Y. (yokoyama@wustl.edu).

LETTERS

The origin of the naked grains of maize

Huai Wang^{1*}, Tina Nussbaum-Wagler^{1*}, Bailin Li², Qiong Zhao¹, Yves Vigouroux^{1†}, Marianna Fallert², Kirsten Bomblies¹, Lewis Lukens³ & John F. Doebley¹

The most critical step in maize (*Zea mays* ssp. *mays*) domestication was the liberation of the kernel from the hardened, protective casing that envelops the kernel in the maize progenitor, teosinte¹. This evolutionary step exposed the kernel on the surface of the ear, such that it could readily be used by humans as a food source. Here we show that this key event in maize domestication is controlled by a single gene (*teosinte glume architecture* or *tga1*), belonging to the SBP-domain family² of transcriptional regulators. The factor controlling the phenotypic difference between maize and teosinte maps to a 1-kilobase region, within which maize and teosinte show only seven fixed differences in their DNA sequences. One of these differences encodes a non-conservative amino acid substitution and may affect protein function, and the other six differences potentially affect gene regulation. Molecular evolution analyses show that this region was the target of selection during maize domestication. Our results demonstrate that modest genetic changes in single genes can induce dramatic changes in phenotype during domestication and evolution.

The origin of the maize ear has been considered one of the greatest mysteries in both crop domestication³ and plant evolution⁴. Although a wealth of botanical and genetic information has identified the wild Mexican grass teosinte (*Zea mays* ssp. *parviglumis*) as the direct progenitor of maize, the profound differences in the structure of the maize and teosinte female inflorescences (ears) have challenged the formulation of a compelling model for the developmental and genetic steps involved in this evolutionary transition³. At the heart of the problem is the fact that teosinte kernels are tightly encased in structures called cupulate fruitcases, whereas maize kernels are borne uncovered on the surface of the ear (Fig. 1a, b). The strength with which the fruitcase envelops the teosinte kernel and the stony nature of this casing far exceed the relatively flimsy and loosely bound chaff that surrounds the kernels of the ancestors of the other domesticated cereals. Indeed, the stony fruitcase of teosinte had been considered such an obstacle to the use of teosinte as a grain that teosinte was dismissed by some as a possible progenitor of maize⁵. It was argued that the genetic steps to free the grain from this casing and thereby convert teosinte into a useful crop were too complex to have arisen under domestication.

Each of the five to twelve cupulate fruitcases in a teosinte ear is formed from an invaginated internode (cupule) within which the kernel sits, and a glume that covers the opening of the cupule such that the kernel is completely hidden from view (Fig. 1b, d). When mature, the teosinte ear disarticulates into the individual fruitcases, each of which contains one kernel. The fruitcase functions to protect the kernel from predation, and passes unscathed through the digestive tracks of animals, providing a means of biotic seed dispersal⁶. At maturity, teosinte fruitcases are heavily lignified and the epidermal cells are filled with silica, giving the fruitcase a stony appearance⁷. Cupules and glumes are present in maize, but reduced

in size relative to the kernel such that they do not surround the kernel. In maize, these organs form the central cob of the ear to which the kernels are attached (Fig. 1a). Maize glumes are less lignified and contain less silica than their teosinte counterparts⁷. Thus, maize domestication involved a change in ear development such that the cupules and glumes form the internal axis of the ear, rather than casings around the kernels. In a sense, maize domestication involved turning the teosinte ear inside out.

Genetic control of the differences in fruitcase/ear structure between maize and teosinte was previously shown to involve a single quantitative trait locus (QTL) of large effect plus several quantitative trait loci of smaller effect⁸. The large effect QTL segregates as a single mendelian locus in an isogenic background, and has been designated *tga1* (ref. 1). However, whether *tga1* represents a single gene or a complex locus consisting of multiple linked genes remained unknown. In the teosinte genetic background, the maize allele (*Tga1-maize*) causes a reduction in internode invagination such that the cupule is too small to house the teosinte kernel, which becomes exposed on the surface of the teosinte ear (Fig. 1c, e). In maize background, the teosinte allele (*tga1-teosinte1*) causes an enlargement of the cupule and glume (Fig. 1f, g). Moreover, the teosinte allele causes the epidermal cells of the cupule and glume to be filled with silica, a feature that accounts for the hardness of the teosinte cupulate fruitcase⁷. The teosinte allele also produces a thicker layer of lignified cells in the glume. The multiple effects on cupule development, three-dimensional growth of the glume, lignification and silica deposition suggest that *tga1* acts as a regulatory gene at the head of a developmental cascade⁷.

We used the maize genetic map, physical maps for maize inbreds B73 and Mo17, and the rice physical map to develop a set of molecular markers near *tga1* (see Supplementary Information). Starting with marker np1316 on maize chromosome 4 (ref. 1), which is tightly linked to *tga1*, we screened maize bacterial artificial chromosome (BAC) libraries and identified a BAC contig near *tga1*. We used BAC end and other sequences from this contig to BLAST the rice genome, and identified a region on rice chromosome 8 that is collinear with the region near *tga1* on maize chromosome 4. Subsequent BLAST searches using the collinear rice sequence identified a second maize contig near *tga1*. DNA sequence analysis revealed that the two maize contigs overlap, enabling us to assemble a single supercontig of ~1.5 megabases. We used markers within this supercontig and determined that *tga1* was located within the supercontig (see Supplementary Methods).

To fine-map *tga1*, we screened 3,106 F₂ plants segregating for *tga1* with markers b91.k20 and umc1511, and scored the plants for the *tga1* trait (Fig. 2a). Marker b91.k20 is located at one end of the supercontig, and umc1511 is located off the other end of the supercontig. Plants containing crossovers between these two markers were assayed for marker be25.a15, which is at the opposite end of the

¹Laboratory of Genetics, University of Wisconsin, Madison, Wisconsin 53706, USA. ²DuPont Crop Genetics, Wilmington, Delaware 19880-353, USA. ³Department of Plant Agriculture, Crop Science Building, University of Guelph, Guelph, Ontario N1G 2W1, Canada. [†]Present address: Institut de Recherche pour le Développement, Montpellier 34730, France.

*These authors contributed equally to this work.

supercontig from b91.k20. These data confirmed that *tga1* lies between b91.k20 and be25.a15. Markers from within the supercontig were placed on our genetic and physical maps, enabling us to map *tga1* to BAC c126f15. Markers within this BAC allowed us to map *tga1* to a ~6-kb segment between bnl252 and bm22.7. As this ~6-kb region encompassed seven additional recombination events, we were able to further map the element that controls the difference between the maize and teosinte phenotypes to a 1,042-bp segment between single nucleotide polymorphisms (SNPs) –1024 and +18.

BLAST searches using the 6-kb region at *tga1* revealed that it has homology to SBP (squamosa-promoter binding protein) transcriptional regulators². Although there are no maize expressed-sequence-tags (ESTs) that closely match this region, we identified matching ESTs from other cereals, including one from rice (GenBank AK109469) that maps to the collinear region on rice chromosome 8 discussed above. On the basis of their homologous sequences and collinear chromosomal locations, AK109469 is probably the rice orthologue of *tga1*. We aligned this rice EST with the maize genomic sequence and identified three exons in maize that match those of rice with putatively conserved start and stop codons (Fig. 2b). We designed primers on the basis of the maize sequence, corresponding to the 3' and 5' non-translated regions of the rice EST. Polymerase chain reaction with reverse transcription (RT-PCR) using total RNA isolated from maize ears gave a single product of the expected size, on the basis of the rice EST. The RT-PCR product was sequenced, confirming that it matched the maize genomic sequence and that it contains two introns in the same positions as in rice (Fig. 2b). The open reading frame encodes a putative protein of 432 amino acids

that is 58% identical and 66% similar to the rice EST (Fig. 2b).

We used another *tga1* allele (*tga1-ems1*) that was generated by ethyl methanesulfonate mutagenesis of maize line W22 to confirm that the SBP-domain gene just described is *tga1* (see Supplementary Methods). Plants homozygous for *tga1-ems1* (Fig. 1h) match the phenotype of the teosinte allele, although phenotypic expression of *tga1-ems1* seems more environmentally labile than *tga1-teosinte1*. DNA sequence analysis of the SPB gene for the *tga1-ems1* stock revealed that it differs from its parental (W22) allele by a non-conservative amino acid substitution of a phenylalanine for a leucine at position 5 (Fig. 2b). This mutation in the *tga1-ems1* allele confirms our conclusion from positional cloning that *tga1* is the SBP gene, and demonstrates that a single amino acid substitution is sufficient to confer the difference between the maize and teosinte phenotypes.

The functional difference between the maize and teosinte alleles of *tga1* could result from differences in gene expression or the *tga1* protein. To investigate the former possibility, we used a combination of northern blots, *in situ* hybridization and real-time PCR with maize inbred W22, an isogenic version of W22 (W22:*tga1*) that carries a teosinte allele at *tga1*, and teosinte itself. Northern blot analysis showed that *tga1* is expressed relatively strongly in immature ears and weakly in husks, but *tga1* expression was not detected in the other tissues examined (Fig. 3a). The messenger RNA levels for immature ears with the maize versus teosinte alleles seemed equivalent. To confirm equivalent expression of the maize and teosinte alleles, we used real-time PCR, which showed that relative *tga1* mRNA levels for the maize (0.79 ± 0.06 , mean $\pm$ s.e.m.) and teosinte (0.85 ± 0.06) alleles were indeed statistically equivalent (t -test = 0.65, $P = 0.37$).

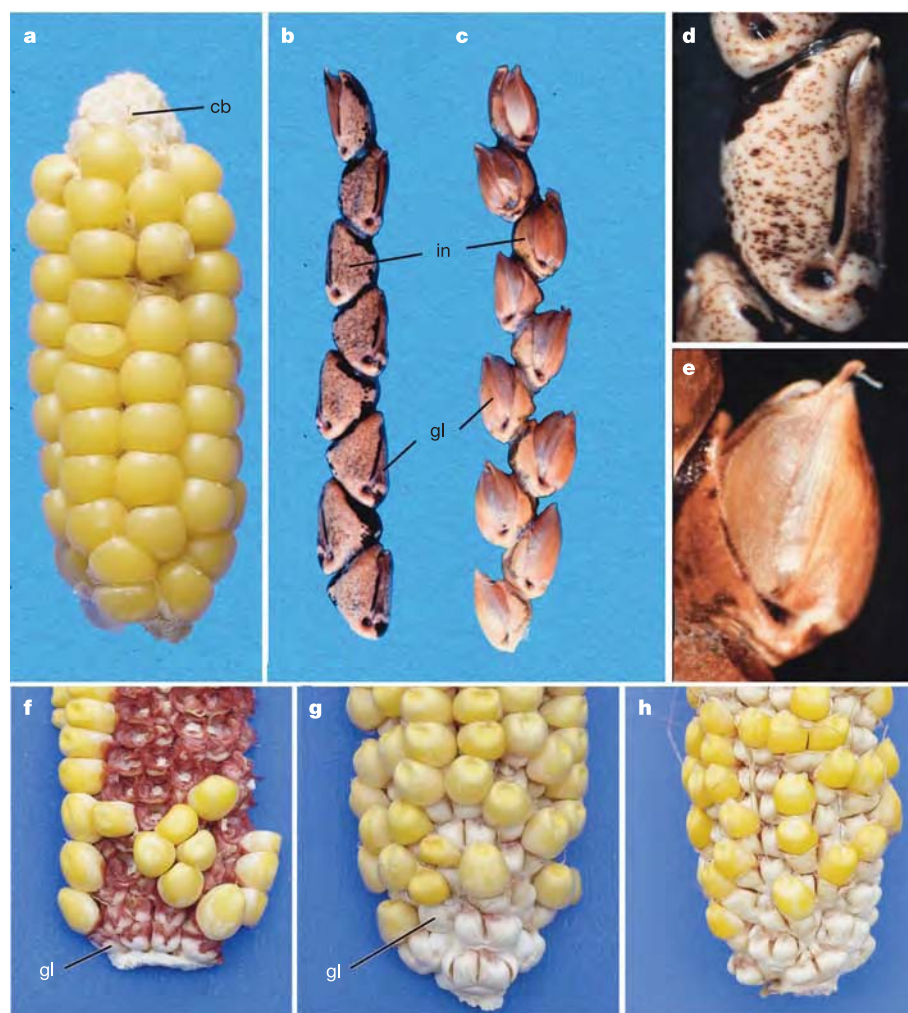


Figure 1 | Phenotypes. **a**, Maize ear showing the cob (cb) exposed at top. **b**, Teosinte ear with the rachis internode (in) and glume (gl) labelled. **c**, Teosinte ear from a plant with a maize allele of *tga1* introgressed into it. **d**, Close-up view of a single teosinte fruitcase. **e**, Close-up view of a fruitcase from teosinte plant with a maize allele of *tga1* introgressed into it. **f**, Ear of maize inbred W22 (*Tga1-maize* allele) with the cob exposed, showing the small white glumes at the base. **g**, Ear of maize inbred W22:*tga1*, which carries the teosinte allele, showing enlarged (white) glumes. **h**, Ear of maize inbred W22 carrying the *tga1-ems1* allele, showing enlarged glumes. For higher magnification images of **f–h**, see Supplementary Information.

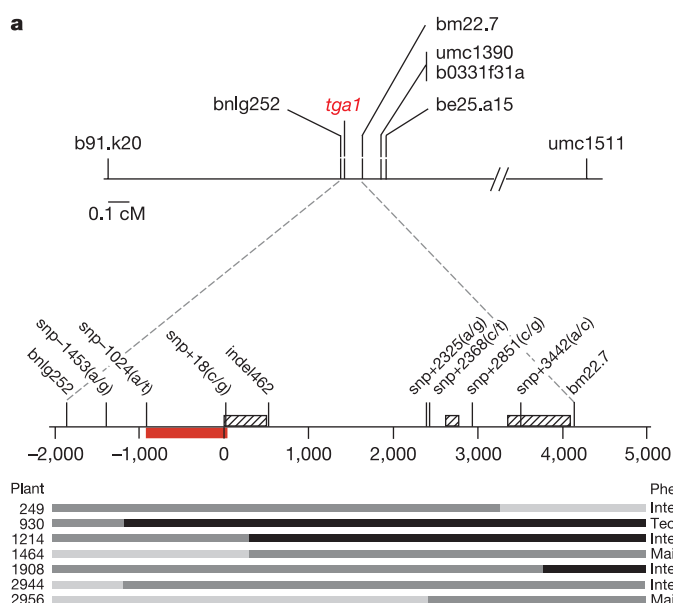
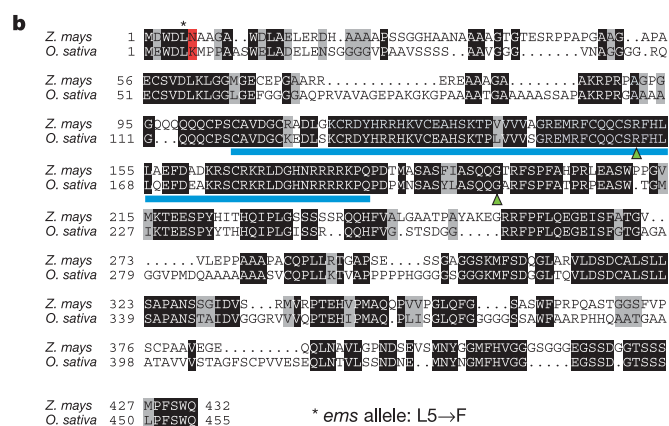


Figure 2 | The *tga1* locus. **a**, Map of the chromosomal segment at *tga1* with the marker loci used for positional cloning. Hatched boxes indicate the three exons of *tga1*. The red box is the 1,042-bp segment to which the factor controlling the difference between maize and teosinte was mapped. Seven of the 3,106 plants with crossovers in *tga1* are shown, together with their phenotypes and genotypes (light grey bar, homozygous maize; medium grey bar, heterozygous; black bar, homozygous teosinte). For details on marker



loci see Supplementary Information. **b**, The *tga1* protein sequence aligned with its rice counterpart (*O. sativa*). Background for identical amino acids is black, grey for similar ones. Blue bar indicates the SBP domain, green triangles indicate the intron positions. The fixed amino acid difference between teosinte and maize (K → N) is highlighted in red. The position of the L → F amino acid change for the *tga1-ems1* allele is marked with an asterisk.

We also compared the expression patterns of the maize and teosinte alleles using *in situ* hybridization. In the maize inbred W22, *tga1* is expressed in the inflorescence meristem of the developing ear and the spikelet pair primordia (Fig. 4a). It is also expressed in the spikelet primordia, including the glume primordia (Fig. 4e). In older spikelet primordia, *tga1* expression is seen in glumes, lemma, lower floret and other floral organs (Fig. 4i). We observed weak but distinct expression throughout the immature ear when hybridization was compared between the antisense probe (Fig. 4a, e, i) and sense controls (Fig. 4d, h, l). We also observed a band of *tga1* expression at the adaxial junction of the spikelet and the inflorescence axis, the region of the inflorescence that develops into the cupule (Fig. 4i).

When we compared the pattern of *tga1* expression seen in W22 to that in W22:*tga1* or teosinte itself, we found no clear differences. In W22:*tga1*, expression is observed in the inflorescence meristem, spikelet pair meristems, glumes, cupule forming region and other floral organs (Fig. 4b, f, j). Similarly, in teosinte, we observed the same spatial pattern of expression (Fig. 4c, g, k). Teosinte also shows *tga1* expression in the husk leaf that subtends the ear (Fig. 4c), consistent with the weak expression seen in maize husks (Fig. 3a). Overall, we do not see any quantitative or qualitative differences in *tga1* expression between the isogenic lines (W22 and W22:*tga1*) or between the maize inbred W22 and teosinte itself. Our expression analyses cannot rule out the existence of some complex or subtle difference in expression, such as prolonged expression throughout later development for one of the genotypes. However, the absence of any discernible differences suggests that differences between the maize and teosinte proteins may be critical to phenotype.

Our genetic analyses narrow the location of the causative site for the functional difference between maize and teosinte to the 1,042-bp segment described above. DNA sequence analysis of this 1,042-bp segment using 16 diverse maize and 12 teosinte individuals identified seven fixed differences between maize and teosinte. Six of these seven are single base-pair polymorphisms that lie just 5' of the coding sequence and potentially affect *tga1* expression. The seventh difference encodes an amino acid substitution of lysine (K) in teosinte to

asparagine (N) in maize at position 6 (Fig. 2b). Several observations suggest that this amino acid substitution is the causative site. First, we observed no differences in the level of *tga1* mRNA accumulation between the isogenic lines for the maize and teosinte alleles. Second, we observed no differences in the pattern of *tga1* expression between

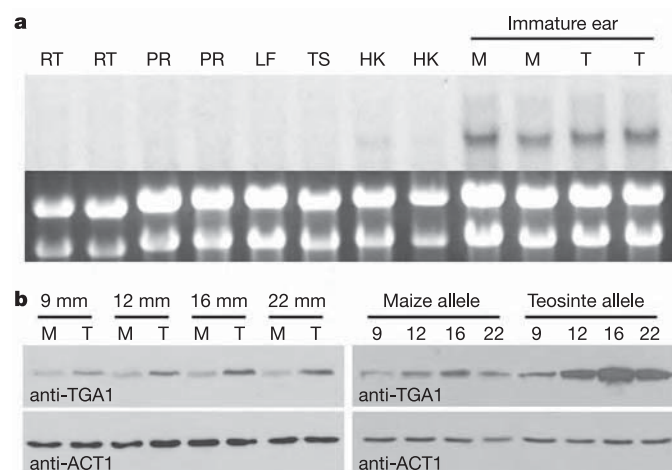


Figure 3 | Molecular analysis of *tga1*. **a**, Top panel is a northern blot showing that *tga1* is expressed in immature ears of W22 maize (M) and W22:*tga1* (T), which carries the teosinte allele. Weak expression is also visible in the husks (HK). No expression of *tga1* is seen in root (RT), prop root (PR), unexpanded leaf (LF) or immature tassel (TS). Lower panel is the ethidium-bromide-stained gel with the visible ribosomal RNAs confirming approximately equal loading of total RNA in each lane. **b**, Western blots showing that the protein encoded by the teosinte allele of *tga1* accumulates at a higher level than the protein of the maize allele. Blots probed with either anti-TGA1 or anti-actin (ACT1, control) antibodies. Protein was extracted from individual ears of W22 maize (M) carrying a maize allele and W22:*tga1* (T) carrying a teosinte allele. Immature ears were staged by their lengths, which are indicated in millimetres.

these same isogenic lines, and teosinte itself shows a parallel pattern of *tga1* expression to that observed in maize. Third, the lysine residue observed in teosinte at position 6 is conserved between rice, wheat (GenBank CK207354) and teosinte, suggesting that it is important to protein function or stability. Fourth, the *tga1-ems1* mutant allele that we recovered alters the amino acid adjacent to the maize/teosinte amino acid difference (Fig. 2b), indicating that this region of the protein is critical and that an amino acid change here is sufficient to distinguish the maize and teosinte phenotypes.

We used western blotting to test whether there is a difference in *tga1* protein abundance associated with the maize and teosinte alleles, and observed that the protein encoded by the teosinte allele is more abundant over a range of developmental stages (Fig. 3b). Given this result, a difference in protein abundance might underlie the phenotypic differences. The K → N substitution might alter protein stability, or it might affect translation efficiency or protein function. One could argue that the difference in protein level is caused by one or more of the six promoter SNPs and that these SNPs are the causative site(s). However, given that the *tga1-teosinte1* and *tga1-ems1* alleles share amino acid changes at adjacent sites, and that these alleles result in nearly the same phenotype, we believe that the K → N substitution is the more probable candidate for the causative site.

If *tga1* were the target of selection during maize domestication, then the signature of past selection may be evident in its level of DNA sequence polymorphisms. We analysed sequence variation across *tga1* for diverse samples of maize and teosinte. Three expectations of past selection were assessed. First, we estimated the ratio of nucleotide diversity (π) in maize to that in teosinte. Selection during maize domestication will reduce this ratio⁹. Second, we calculated Tajima's *D*-statistic¹⁰, which measures whether there is an excess of low frequency polymorphic sites. Such an excess is expected in the wake of a recent selective sweep and will cause a negative *D*-statistic. Third, we applied the HKA (Hudson-Kreitman-Aguadé) test¹¹, which assesses the ratio of diversity in the focal species (maize) to divergence from an outgroup (*Z. diploperennis*) for a target gene (*tga1*) relative to one or more control (neutral) genes.

The *tga1* promoter region shows strong evidence of a past selective

sweep, with a significant Tajima's *D*-statistic and a highly significant HKA test (Fig. 5a). Furthermore, the ratio of $\pi_{\text{maize}}/\pi_{\text{teosinte}}$ indicates that maize possesses only 5% of the diversity found in *parviglumis* teosinte, which is far below the 60–80% level observed for neutral genes^{9,12} and nearly as low as the level observed in the 5' regulatory region of *tb1*, another maize domestication gene¹³. The first exon shows modest evidence for past selection, with a significant HKA test, a $\pi_{\text{maize}}/\pi_{\text{teosinte}}$ ratio of 27% and a negative but not statistically significant *D*-statistic. In contrast, both exons 2 and 3 show neutral patterns of sequence diversity, indicating that the effect of the selective sweep does not extend across the entire gene (Fig. 5a). The stronger evidence for selection in the promoter compared with exon 1 seems inconsistent with the inference that the causative site is the K → N substitution in exon 1. However, the putative selected site is essential at the promoter–exon 1 border, and thus promoter diversity should be affected by selection on this site. Moreover, the promoter has high diversity in teosinte, whereas exon 1 has low diversity in teosinte. This difference means that there is more statistical power to obtain a significant test in the promoter than in exon 1.

We also applied two approaches that model a selective sweep using the coalescent. First, we used the method of Kim and Stephan¹⁴, which calculates the likelihood ratio for a selective sweep versus a neutral model and provides an estimate of the strength of selection ($2N_s$). This analysis indicated that $2N_s = 9,232$, a value which would be observed very rarely under neutral evolution ($P < 0.0001$). Second, we used the method of Przeworski¹⁵, which provides an estimate of the time since fixation of a beneficial allele ($T = T_{\text{gen}}/4N$), which will be < 0.2 in the case of a recent selective sweep. For *tga1*, the mode of the distribution of simulated values of T is ~ 0.021 , and over 99.7% of the simulation values are ≤ 0.2 , providing strong support for a recent selective sweep. This latter method also provides a joint posterior distribution of the selection coefficient (s) and the time in generations since the fixation of the favoured allele ($T_{\text{gen}} = 4NT$). When the population size (N) is set to 100,000, this distribution suggests a selection coefficient of 3–4% and a time since fixation of the favoured allele of $\sim 10,000$ yr for *tga1* (Fig. 5b).

Our results demonstrate that complex differences between the cob

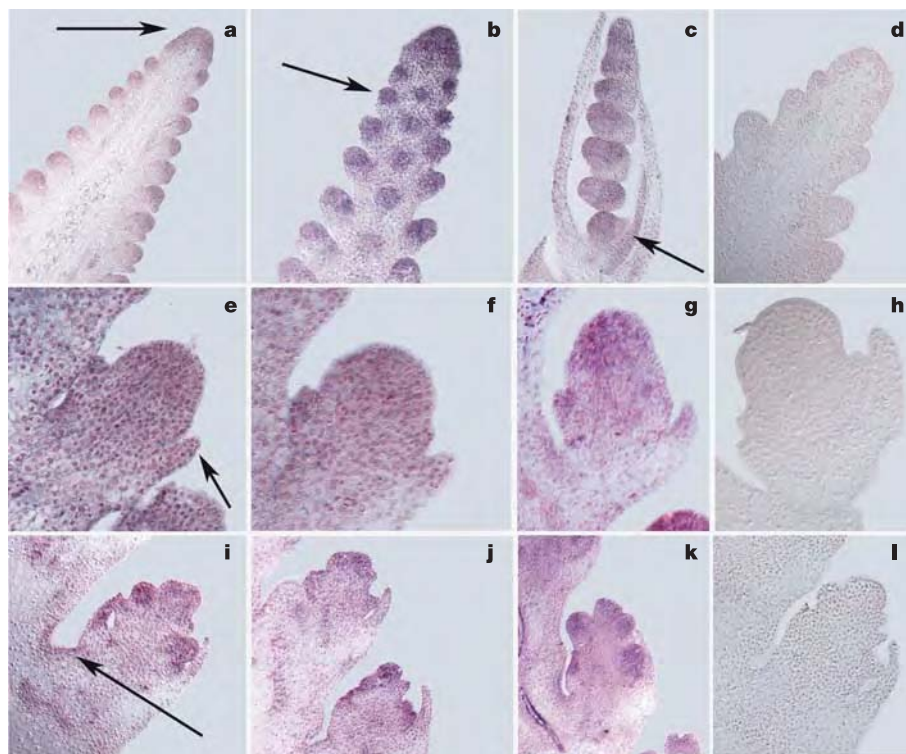


Figure 4 | Tissue *in situ* hybridizations.

a–d, Immature ears. **e–h**, Young spikelet primordia. **i–l**, Older spikelet primordia with distinct floral organ primordia. *tga1* antisense probe was applied to W22 maize (**a**, **e**, **i**), W22:*tga1* (**b**, **f**, **j**) and teosinte (**c**, **g**, **k**), and a *tga1* sense probe was applied to W22 as a control (**d**, **h**, **l**). Arrows show *tga1* expression in the inflorescence meristem (**a**), spikelet-pair primordium (**b**), teosinte husk (**c**), glume (**e**) and cupule-forming region (**i**).

that bears the naked grains of maize and the hardened fruitcases that surround the kernels of teosinte are regulated by a single gene, *tga1*, and it is possible that just a single amino acid change within this gene is responsible for these differences. Exactly how *tga1* regulates ear development remains to be determined. One interpretation is that like many other developmental genes in plants, *tga1* regulates organ identity. The glumes and internodes of the teosinte ear are exceptional among grasses for their degree of lignification and the presence of silica in all epidermal cells. The glumes of the maize ear are more leaf-like and closer in morphology to the glumes of other grasses⁷. In teosinte, therefore, *tga1* may both activate the developmental processes to produce the hardened glumes and invaginated internodes of teosinte, as well as repress a default developmental programme for the leaf-like identity of glumes and solid internodes found in other grasses.

The domestication of maize has been a topic long surrounded by controversy. Mangelsdorf and Reeves⁵ argued that teosinte could not be the progenitor of maize because the morphological differences between maize and teosinte were so vast that the underlying genetic steps could not have arisen during the few thousand years in which maize was domesticated. Thus, they argued that maize evolved in nature over a long evolutionary period. Their assertions were challenged by Beadle¹⁶, who contended that mutations of large effect in a small number of genes would be sufficient to convert teosinte into a useful grain crop. Our results with *tga1* confirm Beadle's

interpretation that simple mutations in single genes contributed major morphological steps in maize domestication.

Although progress has been made in identifying genes involved in crop domestication, identifying the specific polymorphisms controlling domesticated phenotypes has proved elusive. Regulatory changes in the *fw2.2* gene appear to control a major portion of the difference in fruit weight between wild and cultivated tomatoes¹⁷, but the specific regulatory elements involved have not yet been pinpointed. Similarly, regulatory changes in the *tb1* gene have been suggested to control the difference in plant architecture between maize and teosinte^{13,18}, but again the specific regulatory changes remain unknown. For *tga1*, we have identified a set of seven SNPs, one or some combination of which represent the causative site(s). On the basis of the available evidence, our preferred hypothesis is that the K → N substitution controls the phenotypic difference between maize and teosinte.

A feature of all domestication genes that have been identified to date is that the 'cultivated' allele is also found at moderate frequencies in the wild progenitor. This is true of *fw2.2* in tomato¹⁹, *tb1* in maize¹³, and *BoCal* in broccoli²⁰. Thus, the evolution of fruit weight in tomato, plant architecture in maize, and inflorescence structure in broccoli all fit a model of human selection acting upon standing allelic variation that preexisted in the wild progenitor. In contrast, *tga1* may represent a case of new mutation (or at least a rare variant), as we failed to find the maize allele in teosinte by our sequence diversity analysis.

Charles Darwin²¹ used evolution under artificial selection (domestication) as a model for evolution under natural selection. Domestication genes such as *tga1* can provide appropriate models for genes that control key innovations that differentiate natural species. In this context, several features of *tga1* are notable. First, *tga1* represents a single major gene that controls a profound morphological step in maize evolution, and thereby provides support for the view that major gene changes can and do contribute to the origin of evolutionary novelties²². Second, the argument that such gene changes are unlikely because they are typically accompanied by deleterious pleiotropic effects²³ does not apply to *tga1*. Rather, the teosinte allele provides for a fully formed cupulate fruitcase that protects the kernel in the natural environment, whereas the maize allele allows an exposed kernel to be used as food by humans as part of a mutual relationship. The lack of deleterious pleiotropic effects associated with *tga1* can be readily explained by the modular nature of maize (plant) development; *tga1* is expressed in the ear, but not elsewhere in the plant. Thus, changes in *tga1* will not have pleiotropic effects on roots, leaves, stalk or tassel, facilitating ear-specific modifications in function. Finally, the nature of the functional differences in *tga1* between maize and teosinte are simple, no more than 7-bp substitutions and perhaps just a single amino acid change. Our observations strengthen the argument that large phenotypic effects can be caused by very simple molecular changes during domestication or evolution.

METHODS

Genetic mapping. A W22 × W22:*tga1* F₂ fine-mapping population of 3,106 plants that segregated for *Tga1*-maize versus *tga1*-teosinte1 was generated by self-pollinating F₁ hybrids of the parental stocks. This population was screened with markers b91.k20 and umc1511, which flank *tga1*. Plants with crossovers between these two markers were screened with additional markers located between b91.k20 and umc1511. Phenotypes of F₂ plants with crossovers between markers b91.k20 and be25.a15 were scored by visual inspection of mature ears. Although *Tga1*-maize is visually dominant to *tga1*-teosinte1, the heterozygous class is distinct and intermediate between the two homozygous classes⁷, and we were thus able to fully classify the plants into three phenotypic classes. For a complete list of the marker loci, primer sequences, PCR conditions and gel conditions, see the Supplementary Information.

Nucleic acid analyses. After genetic mapping demonstrated that *tga1* lies within the BAC c126f15, we shotgun-sequenced it to 7 × coverage (see Supplementary Information). This work produced a final assembly of ~169 kb in length

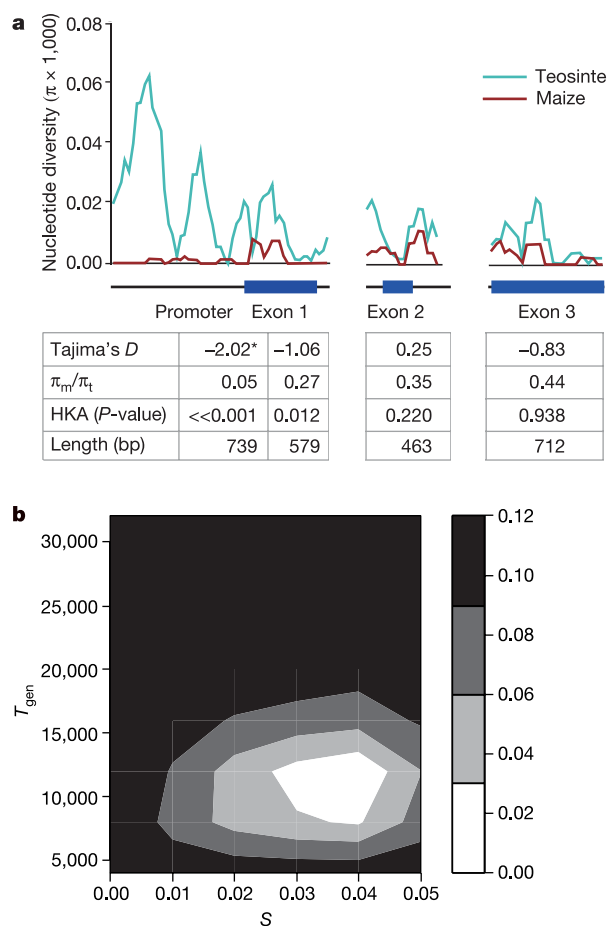


Figure 5 | Molecular evolution. **a**, Nucleotide diversity (π) in maize and teosinte for *tga1*. Tajima's *D*-statistic, HKA tests for non-neutral evolution, and the ratio of π in maize (π_m) to π in teosinte (π_t) are shown. Asterisk, $P < 0.05$. **b**, Joint distribution of the posterior probabilities (scale from white to black) for the time in generations since fixation of the maize allele (T_{gen}) and the selection coefficient *s*.

(GenBank AY883559). BLAST (<http://www.ncbi.nlm.nih.gov/>) searches identified a match between the ~6-kb region between bnlg252 and bm22.7 to which *tga1* was mapped and the rice EST AK109469. We used this rice EST and the corresponding maize genomic sequence as guides to predict the location of the intron–exon boundaries of *tga1*. We confirmed that the predicted maize gene is expressed by RT–PCR. One microgram of total RNA from young W22 ears was reverse-transcribed using Superscript III (Invitrogen) and a *tga1*-specific primer. We then performed PCR with primers that amplify a region from 31 bp 5' of the putative start codon to 99 bp 5' of the predicted stop codon. The PCR product was sequenced to verify gene structure (GenBank AY883560). We isolated and sequenced the exons and 5' non-transcribed regions of *Tga1*-maize, *tga1-teosinte1* and *tga1-ems1* alleles (GenBank AY883561–AY883568) by PCR. For primers, PCR and sequencing conditions, see Supplementary Information.

Gene expression assays. Total RNA was isolated from immature ears (2–3 cm in length) and other tissues using TRI reagent (Molecular Research Center). Northern blots using ~12 µg total RNA of sample per lane were performed using standard procedures (see Supplementary Information). Complementary DNA for real-time PCR was produced using 1 µg of total RNA isolated from 16 immature ears each of W22 and W22:*tga1*. Immature ears were developmentally staged by measuring ear length to ensure that both the W22 and W22:*tga1* genotypes were represented by developmentally equivalent samples (27 ± 0.9 mm and 28 ± 0.9 mm, respectively). RNA was treated with amplification-grade DNaseI (Invitrogen) and then reverse transcribed into cDNA with random hexamers using Taqman Reverse Transcription Reagent Kit (Applied Biosystems). Maize *beta-tubulin2* (GenBank X52879) was selected as an endogenous control for PCR quantification. Two replicates were analysed for each of the 32 samples. For further details about tissue samples, reaction conditions, and primer and probe sequences, see Supplementary Information.

Methods for preparing tissue samples and *in situ* hybridization with digoxigenin-labelled RNA probes followed procedures described elsewhere²⁴. An 891-bp RT–PCR fragment of *tga1* that excluded the SBP-domain and included 80 bp of the 3' untranslated region (UTR) was amplified using PCR and cloned into pGEM-T vector (Promega). Two plasmids with different insert orientations were isolated and linearized using *SpeI*, and then used as templates to generate antisense and sense probes. Probes were synthesized by T7 RNA polymerase using the DIG RNA labelling kit (Roche).

Protein assay. The 3' end of *tga1* excluding the SBP-box was amplified by PCR from a *tga1* RT–cDNA clone (Supplementary Methods). The PCR fragment was cloned into pET151/D-TOPO vector (Invitrogen) and transformed into BL21 codonPlus (DE3)-RPL cells (Stratagene). The histidine-tagged TGA1 (amino acids 181–432) fusion protein was purified using a His-bind purification kit (Novagen). After a further purification by SDS–PAGE gel, the TGA1 fusion protein was used as an antigen to generate a rabbit polyclonal anti-TGA1 antibody (Invitrogen). The specificity of the antibody for TGA1 was tested by western blot analysis.

Protein from immature maize ears (9 mm to 22 mm in length) was extracted using a plant total protein extraction kit (Sigma). Protein concentrations were determined by protein assay using Quick Start Bradford dye reagent (Bio-Rad). For each ear sample, 25 µg of protein was loaded on a 10% SDS–PAGE gel and transferred to an immobilon-P membrane (Millipore) using a mini trans-blot cell (Bio-Rad). The blot was probed with anti-TGA1 primary antibody followed by affinity-purified HRP-conjugated goat anti-rabbit secondary antibody (Kierkegaard & Perry Laboratory). Immune-Star HRP substrate (Bio-Rad) was added and the chemiluminescent signal on the blots was detected with X-ray film. The blots were re-probed with an anti-actin antibody (sc-1616R; Santa Cruz) as a loading control.

Molecular evolution. We sequenced the *tga1* promoter and coding regions for a set of 16 diverse land races of maize and 12 teosinte individuals (*Z. mays* ssp. *parviglumis*) and the outgroup *Zea diploperennis* (GenBank AY883436–AY883558) using PCR primers (see Supplementary Information). PCR products from *Z. diploperennis* were cloned into the TA vector (pCR 2.1-TOPO kit, Invitrogen), and at least four clones were sequenced. Nucleotide diversity (π), Tajima's *D*-statistic¹⁰ and Hudson–Kreitman–Aguadé (HKA) tests¹¹ were calculated in DnaSP Version 4.0 (ref. 25). For the HKA tests, *Z. diploperennis* was used as the outgroup, and *adh1*, *adh2*, *gbl1* and *te1* were used as control loci^{9,12}. The overall HKA *P*-value was obtained by summing the individual χ^2 values for the four control genes.

Received 23 February; accepted 31 May 2005.

1. Dorweiler, J., Stec, A., Kermicle, J. & Doebley, J. *Teosinte glume architecture1*: A genetic locus controlling a key step in maize evolution. *Science* **262**, 233–235 (1993).
2. Klein, J., Saedler, H. & Huijser, P. A new family of DNA binding proteins includes putative transcriptional regulators of the *Antirrhinum majus* floral meristem identity gene *SQUAMOSA*. *Mol. Gen. Genet.* **250**, 7–16 (1996).
3. Doebley, J. The genetics of maize evolution. *Annu. Rev. Genet.* **38**, 37–59 (2004).
4. Maynard Smith, J. Macroevolution. *Nature* **289**, 13–14 (1981).
5. Mangelsdorf, P. C. & Reeves, R. G. The origin of maize. *Proc. Natl Acad. Sci. USA* **24**, 303–312 (1938).
6. Wilkes, H. G. *Teosinte: The Closest Relative of Maize* (Bussey Institution of Harvard Univ., Cambridge, 1967).
7. Dorweiler, J. & Doebley, J. Developmental analysis of *teosinte glume architecture1*: a key locus in the evolution of maize (Poaceae). *Am. J. Bot.* **84**, 1313–1322 (1997).
8. Doebley, J. & Stec, A. Inheritance of the morphological differences between maize and teosinte: comparison of results for two F₂ populations. *Genetics* **134**, 559–570 (1993).
9. Tenaillon, M. I., U'Ren, J., Tenaillon, O. & Gaut, B. S. Selection versus demography: A multilocus investigation of the domestication process in maize. *Mol. Biol. Evol.* **21**, 1214–1225 (2004).
10. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
11. Hudson, R. R., Kreitman, M. & Aguade, M. A. test of neutral molecular evolution based on nucleotide data. *Genetics* **116**, 153–159 (1987).
12. White, S. E. & Doebley, J. F. The molecular evolution of *terminal ear1*, a regulatory gene in the genus *Zea*. *Genetics* **153**, 1455–1462 (1999).
13. Clark, R. M., Linton, E., Messing, J. & Doebley, J. F. Pattern of diversity in the genomic region near the maize domestication gene *tb1*. *Proc. Natl Acad. Sci. USA* **101**, 700–707 (2004).
14. Kim, Y. & Stephan, W. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics* **160**, 765–777 (2002).
15. Przeworski, M. Estimating the time since the fixation of a beneficial allele. *Genetics* **164**, 1667–1676 (2003).
16. Beadle, G. W. Teosinte and the origin of maize. *J. Hered.* **30**, 245–247 (1939).
17. Cong, B., Liu, J. & Tanksley, S. D. Natural alleles at a tomato fruit size quantitative trait locus differ by heterochronic regulatory mutations. *Proc. Natl Acad. Sci. USA* **99**, 13606–13611 (2002).
18. Doebley, J., Stec, A. & Hubbard, L. The evolution of apical dominance in maize. *Nature* **386**, 485–488 (1997).
19. Nesbitt, T. C. & Tanksley, S. D. Comparative sequencing in the genus *Lycopersicon*. Implications for the evolution of fruit size in the domestication of cultivated tomatoes. *Genetics* **162**, 365–379 (2002).
20. Purugganan, M. D., Boyles, A. L. & Suddith, J. I. Variation and selection at the *CAULIFLOWER* floral homeotic gene accompanying the evolution of domesticated *Brassica oleracea*. *Genetics* **155**, 855–862 (2000).
21. Darwin, C. *On The Origin of Species by Means of Natural Selection* (John Murray, London, 1859).
22. Gottlieb, L. D. Genetics and morphological evolution in plants. *Am. Naturalist* **123**, 681–709 (1984).
23. Lande, R. The response to selection on major and minor mutations affecting a metrical trait. *Heredity* **50**, 47–65 (1983).
24. Jackson, D. in *Plant Molecular Pathology: A Practical Approach* (eds Gurr, S. J., McPherson, M. J. & Bowles, D. J.) 163–174 (Oxford Univ. Press, Oxford, 1992).
25. Rozas, J. & Rozas, R. DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* **15**, 174–175 (1999).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Baum and J. Wendel for comments, R. Clark for helpful discussions; E. Ananiev, K. Fengler and M. Morgante for help with screening and identifying BAC contigs; the DuPont sequencing group for BAC sequencing; and P. Rose for assistance with genetic mapping. This research was supported in part by grants from the National Institutes of Health and the USDA Hatch program (to J.F.D.), a Howard Hughes Medical Institute predoctoral fellowship (to K.B.), and an Environmental Protection Agency predoctoral fellowship (to L.L.).

Author Information All sequences have been deposited in GenBank under accession numbers AY883436–AY883568. Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to J.F.D. (jdoebly@wisc.edu).

LETTERS

BRAF^{E600}-associated senescence-like cell cycle arrest of human naevi

Chrysiis Michaloglou^{1*}, Liesbeth C. W. Vredeveld^{1*}, Maria S. Soengas^{3*}, Christophe Denoyelle³, Thomas Kuilman¹, Chantal M. A. M. van der Horst⁴, Donné M. Majoor², Jerry W. Shay⁵, Wolter J. Mooi⁶ & Daniel S. Peeper¹

Most normal mammalian cells have a finite lifespan¹, thought to constitute a protective mechanism against unlimited proliferation^{2–4}. This phenomenon, called senescence, is driven by telomere attrition, which triggers the induction of tumour suppressors including p16^{INK4a} (ref. 5). In cultured cells, senescence can be elicited prematurely by oncogenes⁶; however, whether such oncogene-induced senescence represents a physiological process has long been debated. Human naevi (moles) are benign tumours of melanocytes that frequently harbour oncogenic mutations (predominantly V600E, where valine is substituted for glutamic acid) in BRAF⁷, a protein kinase and downstream effector of Ras. Nonetheless, naevi typically remain in a growth-arrested state for decades and only rarely progress into malignancy (melanoma)^{8–10}. This raises the question of whether naevi undergo BRAF^{V600E}-induced senescence. Here we show that sustained BRAF^{V600E} expression in human melanocytes induces cell cycle arrest, which is accompanied by the induction of both p16^{INK4a} and senescence-associated acidic β -galactosidase (SA- β -Gal) activity, a commonly used senescence marker. Validating these results *in vivo*, congenital naevi are invariably positive for SA- β -Gal, demonstrating the presence of this classical senescence-associated marker in a largely growth-arrested, neoplastic human lesion. In growth-arrested melanocytes, both *in vitro* and *in situ*, we observed a marked mosaic induction of p16^{INK4a}, suggesting that factors other than p16^{INK4a} contribute to protection against BRAF^{V600E}-driven proliferation. Naevi do not appear to suffer from telomere attrition, arguing in favour of an active oncogene-driven senescence process, rather than a loss of replicative potential. Thus, both *in vitro* and *in vivo*, BRAF^{V600E}-expressing melanocytes display classical hallmarks of senescence, suggesting that oncogene-induced senescence represents a genuine protective physiological process.

Melanocytic naevi represent an intriguing human setting where an activated oncogene can co-exist with long-term arrested cells. Naevi are very common, clonal and benign tumours of cutaneous melanocytes¹¹, and frequently harbour the V600E mutation in BRAF⁷ (NCBI gene bank re-named the V599E mutation based on newly available sequence data; accession number NM_004333.2; hereafter referred to as BRAF^{E600}). However, in spite of the oncogenic nature of this mutation^{12,13}, an initial phase of naevus growth is typically followed by a near-complete cessation of proliferative activity, which is maintained for many decades^{8,9,14}. Therefore, it is conceivable that growth arrest of naevi results from oncogene-induced

senescence acting as an effective cellular brake against BRAF^{E600}-mediated oncogenic signalling.

To test this hypothesis, we first determined the effect of the naevus-associated BRAF^{E600} mutant on the proliferative capacity of freshly isolated normal human skin melanocytes. A bicistronic lentiviral vector co-expressing BRAF^{E600} and enhanced green fluorescent protein (eGFP) (to monitor infection efficiency, typically approximately 90%; Fig. 1a) was used to transduce melanocytes. Short-term expression of BRAF^{E600} (3–7 days) led to enhanced melanocyte proliferation that was measured by a moderate but reproducible increase in 5-bromodeoxyuridine (BrdU) incorporation (Fig. 1b and data not shown). However, this effect was transient—sustained expression of BRAF^{E600} resulted in marked cell cycle arrest. In most (roughly 80%) of the transduced melanocytes, this was associated with an intense activity of SA- β -Gal, a marker for senescent or stressed cultured cells as well as for aged tissues *in vivo*^{6,15} (Fig. 1c).

The p16^{INK4a} protein is a major tumour suppressor, often highly expressed in senescent cells *in vitro* and inactivated in a variety of human cancers, including 30–70% of melanomas¹⁶. We found that BRAF^{E600}-expressing melanocytes had elevated levels of p16^{INK4a} (Fig. 1d). Notably, the staining for p16^{INK4a} was heterogeneous, with 25–35% of the BRAF^{E600}-expressing melanocytes showing low or undetectable p16^{INK4a} levels, in spite of the fact that over 95% of these cells were arrested (and still expressed the lentiviral cassette, as evident from eGFP expression).

As oncogene-induced senescence *in vitro* has been best defined in primary diploid fibroblasts, we analysed the cellular response to activated BRAF in this 'reference' context. Consistent with previous results obtained with overexpression of Ras^{V12} or Raf1 (refs 6, 17, 18), retroviral delivery of BRAF^{E600} into two different strains of normal human fibroblasts caused a complete cessation of proliferation, along with induction of p16^{INK4a} (see Supplementary Fig. 1a, b). To exclude the possibility that these effects were caused by supra-physiological levels of BRAF^{E600} expression, we designed a system in which its expression levels could be manipulated with considerable precision. Cells were transduced with a mixture of retrovirus encoding a BRAF^{E600} expression cassette and retrovirus encoding a short hairpin (sh)RNA targeting both endogenous wild type and ectopic BRAF^{E600} (for details, see Fig. 2 legend). This strategy allowed expression of BRAF^{E600} to levels similar to those seen in melanoma cells (see Supplementary Fig. 2a), or to levels close to, or indistinguishable from, endogenous BRAF levels (Fig. 2a, compare lanes 4 and 5).

¹Division of Molecular Genetics and ²Division of Pathology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands. ³Department of Dermatology and Comprehensive Cancer Center, University of Michigan, 1500 E Medical Center Dr. Ann Arbor, Michigan 48109, USA. ⁴Department of Plastic, Reconstructive and Hand Surgery, Academic Medical Centre, PO Box 22660 G4-226, 1100 AZ Amsterdam, The Netherlands. ⁵Department of Cell Biology and Harold Simmons Cancer Center, The University of Texas Southwestern Medical Center, 5323 Harry Hines Boulevard, Dallas, Texas 75390, USA. ⁶Department of Pathology, Free University Medical Centre, De Boelelaan 1117, 1081 HV Amsterdam, The Netherlands.

*These authors contributed equally to this work.

The elevation of $p16^{\text{INK4a}}$ levels was maintained on progressively diminishing $\text{BRAF}^{\text{E600}}$ levels (see Supplementary Fig. 2a). We did not see upregulation of $p14^{\text{ARF}}$ by $\text{BRAF}^{\text{E600}}$ (see Supplementary Fig. 3). Furthermore, low levels of $\text{BRAF}^{\text{E600}}$ caused inhibition of both DNA replication (see Supplementary Fig. 2b) and cellular proliferation (Fig. 2b). This arrest was stably maintained, without any significant escape (see Supplementary Fig. 2c). However, it was bypassed by co-expression of the SV40 large tumour antigen (see Supplementary Fig. 2d), arguing against an aphysiological effect and suggesting that

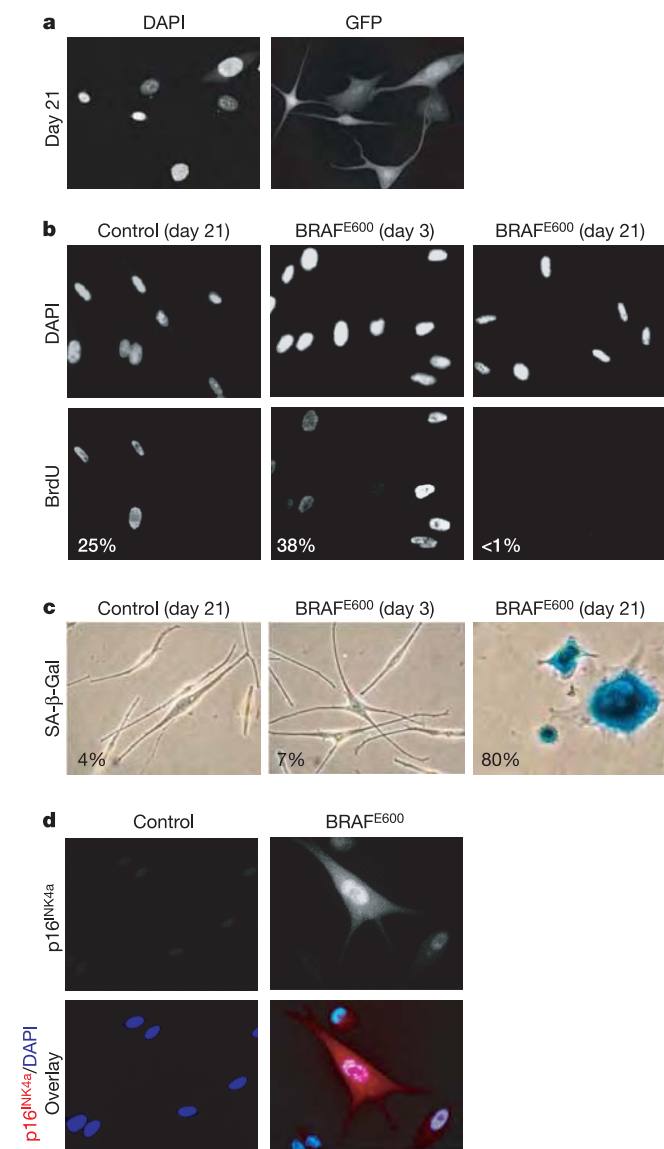


Figure 1 | Sustained expression of $\text{BRAF}^{\text{E600}}$ induces a senescence-like arrest in normal human melanocytes. **a**, Infection efficiency is estimated to be approximately 90% by visualization of eGFP-positive cells. Nuclear staining by DAPI is shown as an indication of total cell number. **b**, Impact of $\text{BRAF}^{\text{E600}}$ on melanocyte proliferation measured as the percentage of positive cells after a 3-h pulse with BrdU. DAPI stain to detect nuclei (top panel) and BrdU-positive cells (bottom panel). Control melanocytes remained proliferative and viable. **c**, SA- β -Gal activity after brief and sustained $\text{BRAF}^{\text{E600}}$ expression. **d**, Heterogeneous levels of $p16^{\text{INK4a}}$ (upper panels and lower panels, red) in $\text{BRAF}^{\text{E600}}$ -transduced melanocytes. DAPI stain is used to detect nuclei (lower panels, blue). Note that at 21 days after infection, <1% of the $\text{BRAF}^{\text{E600}}$ -expressing melanocytes incorporated BrdU, the vast majority showed positive SA- β -Gal activity, whereas a smaller proportion of $\text{BRAF}^{\text{E600}}$ -expressing cells had detectable $p16^{\text{INK4a}}$ expression. Numbers given are representative of three independent experiments.

this arrest depends on cellular tumour suppressors. Furthermore, in the context of fibroblasts defective for p53 and $p16^{\text{INK4a}}$ and expressing SV40 small tumour antigen and human telomerase reverse transcriptase (hTERT), $\text{BRAF}^{\text{E600}}$ could efficiently substitute for Ras^{V12} in the induction of tumours in immunocompromised mice (see Supplementary Table 1).

Recently, overexpression of Ras^{V12} has been shown to cause accumulation of senescence-associated heterochromatic foci (SAHF), concentrated spots of transcriptionally silenced DNA¹⁹. We observed that low levels of $\text{BRAF}^{\text{E600}}$ also induced SAHF (Fig. 2c). This was accompanied by focal accumulation of a specific heterochromatin-associated histone modification, namely methylation of lysine 9 of histone H3 (K9M-H3). Together, these results indicate that low levels of $\text{BRAF}^{\text{E600}}$ induce senescence-like cell cycle arrest in primary human cells.

To investigate whether $p16^{\text{INK4a}}$ upregulation constitutes a protective response to inappropriate mitogenic signalling by $\text{BRAF}^{\text{E600}}$, we created cell lines stably expressing $p16^{\text{INK4a}}$ shRNA. This shRNA was effective, as it suppressed the accumulation of $p16^{\text{INK4a}}$ in various settings (Fig. 2a), caused increased proliferation

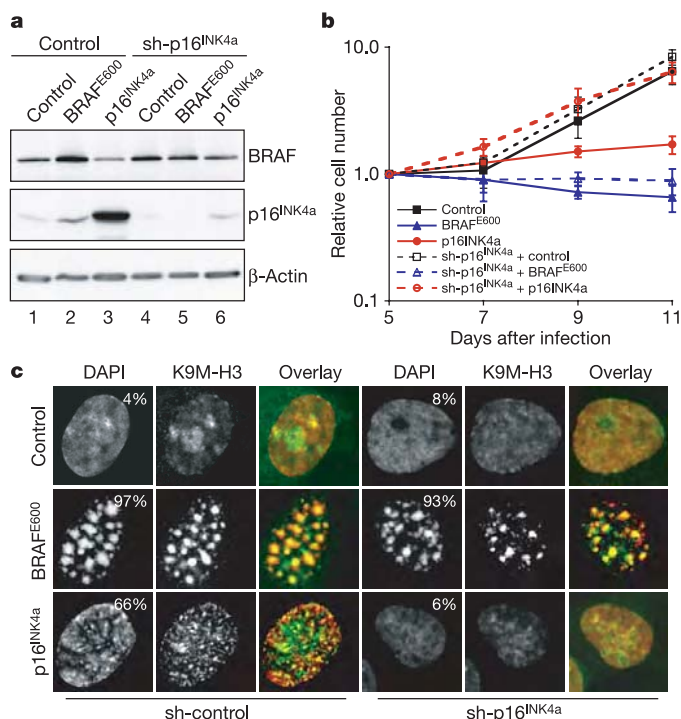


Figure 2 | Physiological levels of $\text{BRAF}^{\text{E600}}$ induce senescence-like arrest in normal human fibroblasts in a $p16^{\text{INK4a}}$ -independent manner. **a–c**, Human BJ_{hTERT} cells stably expressing either control or $p16^{\text{INK4a}}$ shRNA were transduced with a mixture of a retrovirus encoding $\text{BRAF}^{\text{E600}}$ expression and puromycin resistance cassettes, and a retrovirus encoding cassettes for a shRNA targeting both endogenous wild type and ectopic $\text{BRAF}^{\text{E600}}$ and blasticidin resistance (labelled $\text{BRAF}^{\text{E600}}$). Pharmacological selection for both resistance markers was used to create cells that had stably integrated both the $\text{BRAF}^{\text{E600}}$ expression cassette and the shRNA. The latter alone did not affect proliferative potential (data not shown). As a control, cells were transduced with $p16^{\text{INK4a}}$ -encoding retrovirus. **a**, Analysis by western blotting with β -actin used as a loading control. **b**, Analysis by proliferation curves. Shown are the results of three independent experiments performed in duplicate with standard deviations. **c**, DAPI staining to detect SAHF (percentage of SAHF-positive cells is indicated in insert) and immunofluorescence for K9M-H3. Note that whereas $p16^{\text{INK4a}}$ shRNA neutralized senescence induced by ectopic expression of $p16^{\text{INK4a}}$, it failed to bypass $\text{BRAF}^{\text{E600}}$ -induced senescence. Even undetectable levels of $\text{BRAF}^{\text{E600}}$ (**a**, compare lanes 4 and 5) were sufficient to induce growth arrest and SAHF, in the absence of $p16^{\text{INK4a}}$.

(see Supplementary Fig. 4) and abolished the induction of cell cycle arrest by overexpressed $p16^{\text{INK4a}}$ (Fig. 2b; see also Supplementary Fig. 2b, c). However, low levels of $\text{BRAF}^{\text{E600}}$ inhibited proliferation and induced SAHF even on $p16^{\text{INK4a}}$ depletion (Fig. 2b, c; see also Supplementary Fig. 2b, c). Identical observations were made for a $p16^{\text{INK4a}}$ binding-deficient cyclin-dependent kinase 4 mutant ($\text{CDK4}^{\text{R24C}}$; see Supplementary Fig. 2e), for a different fibroblast strain (TIG3; see Supplementary Fig. 2f–h) and for fibroblasts explanted from a homozygous $p16^{\text{INK4a}}$ -deficient 'Leiden' patient²⁰ (see Supplementary Fig. 2i–l). These results suggest that, at least in cultured human fibroblasts, $p16^{\text{INK4a}}$ is upregulated in response to physiological levels of $\text{BRAF}^{\text{E600}}$, but is not strictly required for the induction of cell cycle arrest.

Next, we wished to validate *in vivo* our observation that $\text{BRAF}^{\text{E600}}$ induces senescence-like arrest *in vitro*. Given the high frequency of BRAF mutations in both naevi and melanomas^{7,12}, we looked for hallmarks of senescence in a panel of resection specimens of human naevi. We first confirmed the presence of the $\text{BRAF}^{\text{E600}}$ mutation in eight specimens of our panel of 23 naevi (see Supplementary Fig. 5). We then used paraffin-embedded as well as cryosections of normal human skin and resection specimens of naevi for further analysis. Virtually all melanocytes within these naevi were growth-arrested, as judged by negative immunohistochemical staining for the proliferation marker Ki-67 (Fig. 3a), in agreement with the literature^{8,9}. This was in contrast to epidermal keratinocytes in the same specimens, many of which showed proliferative activity.

As expression of $\text{BRAF}^{\text{E600}}$ in cultured human melanocytes led to cell cycle arrest along with induction of SA- β -Gal activity, we subjected our naevi panel to analysis for this senescence marker.

SA- β -Gal activity has previously been shown to increase in human epidermis as a function of age¹⁵. To avoid detection of age-induced SA- β -Gal activity, we analysed a panel of congenital naevi obtained from patients under one year of age. Indeed, all 23 naevi analysed displayed high levels of SA- β -Gal activity (Fig. 3a–c). This activity was absent from normal skin melanocytes within the same samples (Fig. 3b, c) and in control samples taken from normal skin (Fig. 3a). In agreement with previous data²¹, SA- β -Gal activity was also absent from freshly isolated, cultured primary human melanocytes (Fig. 1c). Terminally differentiated keratinocytes within the epidermis of the same samples also did not show SA- β -Gal activity (Fig. 3b, c, arrowheads), consistent with previous observations made in young (<39 yr of age) donors¹⁵. SA- β -Gal activity was similarly lacking from the cells of skin adnexa, such as the sweat gland (Fig. 3b, c, labelled S), and from the dermal mesenchymal cells between the naevus cell nests. Thus, human naevi, largely growth-arrested neoplastic lesions, are positive for the senescence marker SA- β -Gal.

Studies in mouse models and humans indicate that (epi)genetic inactivation of the $p16^{\text{INK4a}}$ gene is associated with melanomagenesis^{9,16,22,23}. Normal melanocytes, scattered alongside the dermal–epidermal junction and surrounded by keratinocytes, had undetectable levels of $p16^{\text{INK4a}}$ (Fig. 3a). In contrast, naevi invariably contained $p16^{\text{INK4a}}$ -expressing cells, in agreement with previous observations^{9,22}. Of note, we failed to detect any significant upregulation of p53 and p21^{CIP1} in naevi (data not shown). Irrespective of the mutational status of BRAF in the naevi, the percentage of $p16^{\text{INK4a}}$ -positive cells and the intensity of staining per cell was heterogeneous, with a striking mosaic pattern of $p16^{\text{INK4a}}$ immunopositivity seen in most naevi (Fig. 3a; see also Supplementary Fig. 6; data not shown).

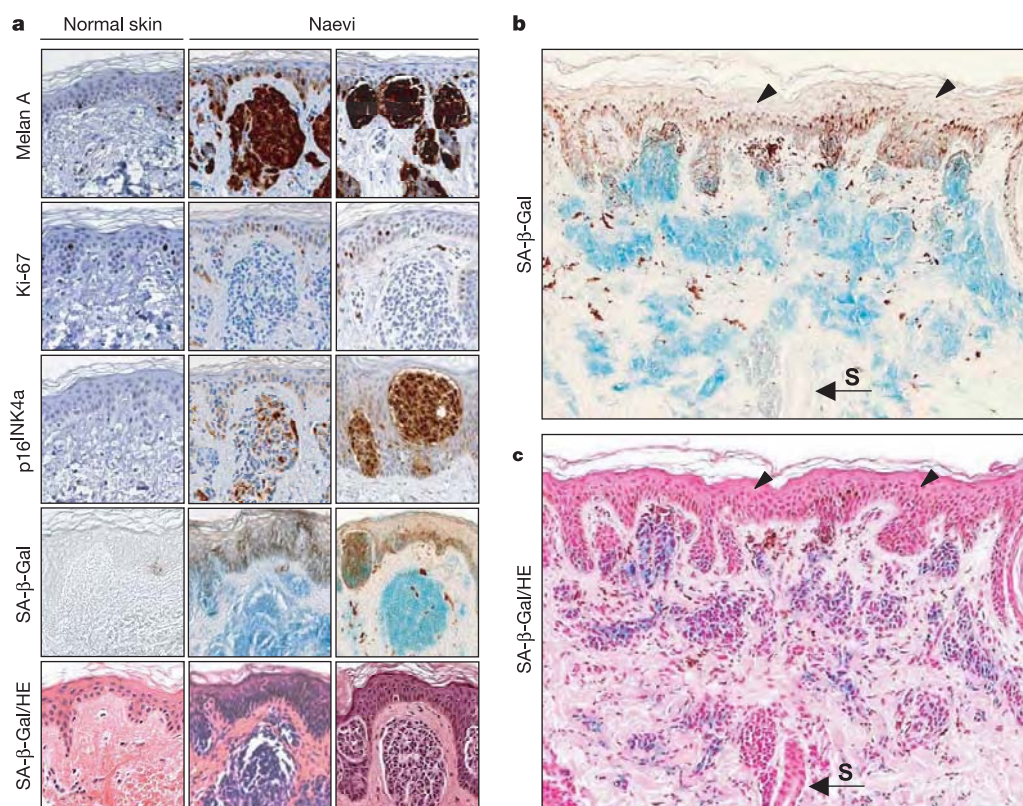


Figure 3 | Human melanocytic naevi display the hallmarks of senescent cells. Paraffin-embedded sections of human naevi and normal skin were subjected to immunohistochemistry with the indicated antibodies. **a**, Melan A (brown) identifies melanocytes, MIB1 (brown) recognizes the proliferation marker Ki-67 and $p16^{\text{INK4a}}$ antibody (brown) detects $p16^{\text{INK4a}}$. **a–c**, Frozen sections of human naevi were subjected to SA- β -Gal staining. The blue staining corresponds exactly to the sites of naevus cell nests. All

cells of the epidermis (mostly keratinocytes, arrowheads) and sweat gland (S), as well as dermal cells between the naevus cell nests, are completely negative for SA- β -Gal (**b** and **c**). Note that the brown staining in these samples comes from pigment released by the naevus cells. Where indicated, haematoxylin and eosin (HE) was used as a counter stain on the consecutive (SA- β -Gal-stained) section to visualize the melanocyte lesions, within the context of the surrounding tissue.

Importantly, all melanocytes within these $p16^{\text{INK4a}}$ -mosaic naevi (whether positive or negative for $p16^{\text{INK4a}}$) were growth arrested (Fig. 3a; see also Supplementary Fig. 6). Indeed, the lack of proliferation of the naevus melanocytes was associated with positive SA- β -Gal staining rather than with $p16^{\text{INK4a}}$ immunoreactivity, thereby closely mimicking $\text{BRAF}^{\text{E600}}$ -expressing melanocytes *in vitro*. As naevi are thought to be clonal¹¹, it is unlikely that a differential BRAF mutation pattern underlies the $p16^{\text{INK4a}}$ mosaicism. The observed $p16^{\text{INK4a}}$ mosaicism is in accordance with the occurrence of growth arrest in naevi of homozygous $p16^{\text{INK4a}}$ -deficient Leiden patients²⁰. However, the naevi of these patients are increased in number and size compared to those of $p16^{\text{INK4a}}$ -proficient individuals. Together with the mosaic $p16^{\text{INK4a}}$ pattern seen in $\text{BRAF}^{\text{E600}}$ -expressing, growth-arrested melanocytes *in vitro*, these observations suggest that $p16^{\text{INK4a}}$ collaborates with other, yet-to-be-identified factors in the establishment of long-term growth arrest of naevi. Whether $p16^{\text{INK4a}}$ contributes to an irreversible arrest that, once established, no longer requires $p16^{\text{INK4a}}$ (as has been suggested previously^{19,24}), or is involved in aspects of melanoma progression other than oncogene-induced senescence remains to be determined. Together these results show that human naevi and $\text{BRAF}^{\text{E600}}$ -expressing melanocytes share three senescence-associated markers: stable growth arrest, heterogeneous induction of $p16^{\text{INK4a}}$ and SA- β -Gal activity.

Although these features support oncogene-induced senescence as

the cause for cell cycle arrest of melanocytic naevi, in principle these senescence markers may have been triggered by telomere attrition, which might have resulted from the initial expansion of melanocytes. Indeed, telomere attrition has been previously proposed as a major factor involved in the cell-cycle arrest of naevi²⁵. This would be consistent with the clinical observation that congenital naevi (that is, arising before birth) can cover large areas of the body surface, in contrast to naevi acquired later in life, which are usually much smaller. To test this possibility, we performed telomere FISH (fluorescent *in situ* hybridization) on congenital naevus resections. As expected, melanoma metastases, stained as controls, displayed much less fluorescence than their surrounding stromal cells, indicating that this method is sufficiently sensitive to detect differences in telomere length *in situ* (Fig. 4a, b; see also Supplementary Fig. 7). In contrast, we did not observe any significant difference in telomere fluorescence when comparing congenital naevi to surrounding tissues, in agreement with previous observations made in acquired and Spitz naevi²⁶. Although we cannot formally rule out the possibility that a single eroded telomere may have triggered the senescence response, one would not expect the other telomeres to remain apparently full length. These results argue in favour of an active oncogene-driven senescence process, rather than senescence triggered by exhaustion of replicative potential resulting from gross telomere attrition.

Our observation that $\text{BRAF}^{\text{E600}}$ initially stimulates moderate melanocyte proliferation supports the hypothesis that it contributes to the initiating events of melanomagenesis. However, our results suggest that, both *in vitro* and *in vivo*, oncogenic BRAF signalling subsequently leads to a growth-inhibitory response, which is associated with the known classical hallmarks of senescence (that is, stable proliferative arrest, an increase in $p16^{\text{INK4a}}$ and the induction of SA- β -Gal activity). Our results therefore provide support for a speculative model previously proposed^{9,27}, in which $\text{BRAF}^{\text{E600}}$ cannot fully transform human melanocytes, but requires additional, cooperating events for tumour development. Supporting this view, zebrafish expressing a $\text{BRAF}^{\text{E600}}$ transgene develop 'fish-naevi', which require a p53-deficient background to progress to invasive melanomas²⁸. Our observations provide evidence of oncogene-induced senescence as a physiological mechanism in humans limiting the progression of premalignant lesions.

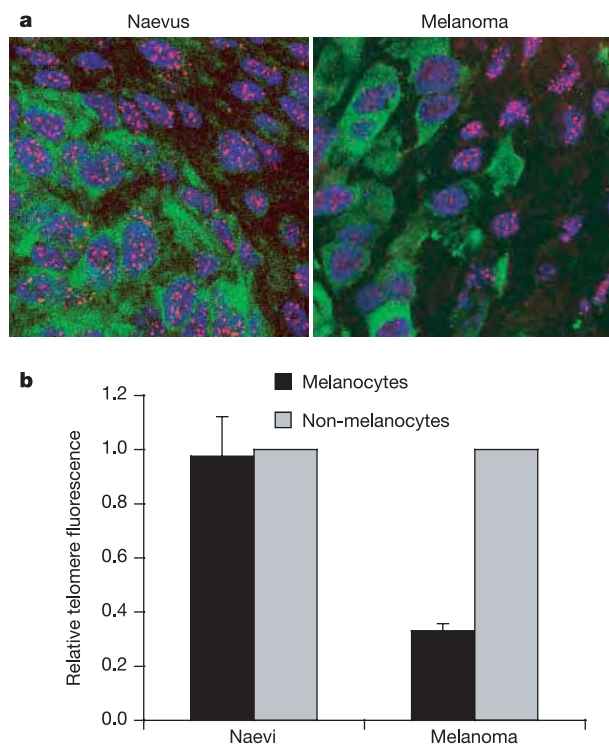


Figure 4 | No apparent telomere loss in naevi. a, Frozen sections of human naevi and melanomas were subjected to telomere FISH (red). A representative sample is shown; six naevi were analysed (five of which harboured the $\text{BRAF}^{\text{E600}}$ mutation). For reference, sections were also stained with the Melan A antibody (detecting melanocytes in naevus and melanoma, green) and DAPI (blue). Note that in the naevus specimen, both the naevus cells (Melan-A positive) and the stromal compartment (Melan-A negative) are FISH positive. In contrast, melanoma tumour cells, which conceivably have undergone telomere attrition, have a much weaker telomere fluorescence signal than their surrounding stromal cells. The channel overlay is shown (individual channels are shown in Supplementary Fig. 7). **b**, Quantification of the data in **a**, plotted as relative telomere fluorescence in six naevi and three melanomas (both indicated as melanocytes) compared to their respective stromal cells (non-melanocytes) with standard deviations.

METHODS

Plasmids. Gene transfer in normal human melanocytes was achieved with two independent lentiviral vectors (FG12-HA-BRAF^{E600}-eGFP and HIV-CS-CG-BRAF^{E600}-puro), with comparable phenotypes. pBabe-puro-BRAF^{E600} was used to transduce fibroblasts.

The $p16^{\text{INK4a}}$ shRNA sequence corresponds to nucleotides 21–41 (GenBank accession number NM_000077). The BRAF shRNA sequence was previously described²⁹.

Cell culture, retroviral transduction, cell cycle analysis and proliferation curves. Normal human melanocytes were isolated from the epidermis of neonatal foreskins. Briefly, foreskins were incubated overnight in trypsinization solution (22.5 mM HEPES, 7.5 mM glucose, 2.25 mM KCl, 100 mM NaCl, 0.75 mM Na₂HPO₄ and 0.17% trypsin). Dermis and epidermis were separated by scraping. The epidermal compartment was incubated for 3–4 d in 254CF medium (Cascade Biologicals) supplemented with 0.1 mM CaCl₂, 2% fetal bovine serum and keratinocyte growth supplement (Cascade Biologicals). Melanocytes were subsequently separated from keratinocytes by differential trypsinization. Purity of the melanocytic preparations was estimated by standard immunostaining for HMB-45. For long-term cultures, melanocytes were propagated in 254CF medium in the presence of 0.2 mM CaCl₂ and melanocyte growth supplement (Cascade Biologicals).

BJ cells expressing hTERT (in pBabeHygro) were grown in DMEM/medium 199 (Gibco) in a 4:1 ratio supplemented with 15% fetal bovine serum (PAA Laboratories), 0.1 mM MEM non-essential amino acids (Gibco), 2 mM glutamine (Gibco), 100 units ml⁻¹ penicillin and 0.1 mg ml⁻¹ streptomycin.

Lentiviral infections were performed using HEK293T cells as producers of viral supernatants. The Phoenix packaging cell line was used for the generation of ecotropic retroviruses, as described³⁰. Amphotropic retroviruses encoding the

ecotropic receptor were generated in HEK293T cells. All infections were carried out as described³⁰.

BrdU labelling was carried out for 3 h in reduced growth factor conditions. Proliferation curves were performed as described^{6,30}.

Human tissue samples. Surgical resection specimens of congenital naevi were obtained from patients in their first year of life. Specimens were fresh frozen and stored for a period of 2 months to 2 years at -70°C before use.

Analysis of SA- β -galactosidase activity. Tissues were fixed in 4% formaldehyde for 2 h, washed with 0.2% NP-40, 0.1% Na-deoxycholate, 2 mM MgCl_2 , 100 mM Na_2HPO_4 , pH 6.0 and stained as described¹⁵. Tissues were post-fixed overnight in 4% formaldehyde and embedded in paraffin. Cultured cells were stained as described¹⁵.

Immunohistochemistry and antibodies. Tissues were fixed in 4% formaldehyde overnight and embedded in paraffin. Paraffin sections were deparaffinized, rehydrated, incubated in 0.1 mM sodium citrate pH 6.0, washed and incubated with peroxidase blocking reagent (S2001; DAKO). The tissue was incubated with the primary antibodies Ab-3 for Melan A (Molecular Probes), MIB-1 for Ki-67 (DAKO) and Ab-4 (16P04) for p16^{INK4a} (Molecular Probes). Secondary antibody used was PowerVision+ (DPVB + 999HRP; ImmunoLogic). Peroxidase activity was detected with Liquid DAB (K3468; DAKO).

The antibodies used for western blotting were F-7 for BRAF (sc-5284; Santa Cruz), JC8 (MS-889; NeoMarkers) for p16^{INK4a}, C-22 for CDK4 (sc-260; Santa Cruz) and AC-74 for β -actin (A5316; Sigma). The antibodies used for immunofluorescence were anti-trimethyl-histone H3 (Lys 9) (07-442; Upstate) and 554070 (BD Pharmingen) for p16^{INK4a}. DAPI staining was used to visualize nuclei.

Telomere FISH. Tissue sections were prepared for FISH analysis with the 2'-deoxyoligonucleotide N3' $\rightarrow$ P5' phosphoramidate probe: 5'-(CCCTAA)₃-TAMRA-3', specific for telomeric sequences. The Melan A antibody Ab-3 was used and also the DNA staining dye DAPI. Images of fluorescein (FITC), TAMRA and DAPI fluorescence were acquired on a digital image microscopy system. The relative telomere length is proportional to the number of hybridized probes. For quantification, the fluorescence signal was determined (using ImageJ software) for all individual telomeres within 12 nuclei per specimen, both for melanocytes (within naevi and melanomas) and for cells within the stromal compartments of both types of lesion, for six independent naevi and three independent melanomas. All measurements have been corrected for background nuclear fluorescence.

Received 19 January; accepted 8 June 2005.

- Hayflick, L. The limited *in vitro* lifetime of human diploid cell strains. *Exp. Cell Res.* **37**, 614–636 (1965).
- Mathon, N. F. & Lloyd, A. C. Cell senescence and cancer. *Nature Rev. Cancer* **1**, 203–213 (2001).
- Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432**, 307–315 (2004).
- Campisi, J. Senescent cells, tumour suppression, and organismal aging: good citizens, bad neighbors. *Cell* **120**, 513–522 (2005).
- Shay, J. W. & Roninson, I. B. Hallmarks of senescence in carcinogenesis and cancer therapy. *Oncogene* **23**, 2919–2933 (2004).
- Serrano, M., Lin, A. W., McCurrach, M. E., Beach, D. & Lowe, S. W. Oncogenic ras provokes premature cell senescence associated with accumulation of p53 and p16^{INK4a}. *Cell* **88**, 593–602 (1997).
- Pollock, P. M. *et al.* High frequency of BRAF mutations in nevi. *Nature Genet.* **33**, 19–20 (2003).
- Kuwata, T., Kitagawa, M. & Kasuga, T. Proliferative activity of primary cutaneous melanocytic tumours. *Virchows Arch. A Pathol. Anat. Histopathol.* **423**, 359–364 (1993).
- Bennett, D. C. Human melanocyte senescence and melanoma susceptibility genes. *Oncogene* **22**, 3063–3069 (2003).
- Chin, L., Merlino, G. & DePinho, R. A. Malignant melanoma: modern black plague and genetic black box. *Genes Dev.* **12**, 3467–3481 (1998).
- Robinson, W. A. *et al.* Human acquired naevi are clonal. *Melanoma Res.* **8**, 499–503 (1998).
- Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417**, 949–954.
- Wellbrock, C. *et al.* V599EB-RAF is an oncogene in melanocytes. *Cancer Res.* **64**, 2338–2342 (2004).
- Mooi, W. J. & Krausz, T. *Biopsy Pathology of Melanocytic Disorders* 56–105 (Chapman & Hall Medical, London, 1992).
- Dimri, G. P. *et al.* A biomarker that identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
- Sharpless, E. & Chin, L. The INK4a/ARF locus and melanoma. *Oncogene* **22**, 3092–3098 (2003).
- Zhu, J., Woods, D., McMahon, M. & Bishop, J. M. Senescence of human fibroblasts induced by oncogenic Raf. *Genes Dev.* **12**, 2997–3007 (1998).
- Lin, A. W. *et al.* Premature senescence involving p53 and p16 is activated in response to constitutive MEK/MAPK mitogenic signalling. *Genes Dev.* **12**, 3008–3019 (1998).
- Narita, M. *et al.* Rb-mediated heterochromatin formation and silencing of E2F target genes during cellular senescence. *Cell* **113**, 703–716 (2003).
- Gruis, N. A. *et al.* Homozygotes for CDKN2 (p16) germline mutation in Dutch familial melanoma kindreds. *Nature Genet.* **10**, 351–353 (1995).
- Bandyopadhyay, D. *et al.* The human melanocyte: a model system to study the complexity of cellular aging and transformation in non-fibroblastic cells. *Exp. Gerontol.* **36**, 1265–1275 (2001).
- Wang, Y. L., Uhara, H., Yamazaki, Y., Nikaido, T. & Saido, T. Immunohistochemical detection of CDK4 and p16^{INK4a} proteins in cutaneous malignant melanoma. *Br. J. Dermatol.* **134**, 269–275 (1996).
- Kamb, A. *et al.* A cell cycle regulator potentially involved in genesis of many tumour types. *Science* **264**, 436–440 (1994).
- Beausejour, C. M. *et al.* Reversal of human cellular senescence: roles of the p53 and p16 pathways. *EMBO J.* **22**, 4212–4222 (2003).
- Bastian, B. C. Understanding the progression of melanocytic neoplasia using genomic analysis: from fields to cancer. *Oncogene* **22**, 3081–3086 (2003).
- Miracco, C. *et al.* Quantitative *in situ* evaluation of telomeres in fluorescence *in situ* hybridization-processed sections of cutaneous melanocytic lesions and correlation with telomerase activity. *Br. J. Dermatol.* **146**, 399–408 (2002).
- Peeper, D. S. & Mooi, W. J. Pathogenesis of melanocytic naevi: growth arrest linked with cellular senescence? *Histopathology* **41**, S139–S143 (2002).
- Patton, E. E. *et al.* BRAF mutations are sufficient to promote nevi formation and cooperate with p53 in the genesis of melanoma. *Curr. Biol.* **15**, 249–254 (2005).
- Hingorani, S. R., Jacobetz, M. A., Robertson, G. P., Herlyn, M. & Tuveson, D. A. Suppression of BRAF(V599E) in human melanoma abrogates transformation. *Cancer Res.* **63**, 5198–5202 (2003).
- Peeper, D. S. *et al.* A functional screen identifies hDRIL1 as an oncogene that rescues RAS-induced senescence. *Nature Cell Biol.* **4**, 148–153 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank D. Atsma, E. Mesman and J. Zevenhoven for help with immunohistochemistry; S. Douma for analytical support; L. Oomen, L. Brocks and J. van Rheenen for help with microscopy; N. Gruis and C. Out for p16^{INK4a}-deficient fibroblasts; L. Zaai and A. van der Wal for help with obtaining congenital naevus specimens; M. Voorhoeve and R. Agami for pRetroSuper, pRetroSuper-Blasticidin and pRetroSuper-GFP; S. Gryaznov for the telomere probe; R. Beijersbergen and M. van Lohuizen for reagents; G. Abou-Rjaily for help with lentiviral infections; P. Krimpenfort and colleagues in the Peeper laboratory for discussions; R. Bernards for support; and M. van Lohuizen and A. Berns for suggestions and reading of the manuscript. M.S.S. is supported by an NIH grant. M.S.S. is a V Foundation for Cancer Research Scholar. L.C.W.V., T.K. and D.S.P. were supported by the Netherlands Organization for Scientific Research (NWO).

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.S.P. (d.peeper@nki.nl) or W.J.M. (wj.mooi@vmc.nl).

Crucial role of p53-dependent cellular senescence in suppression of Pten-deficient tumorigenesis

Zhenbang Chen^{1,2}, Lloyd C. Trotman^{1,2}, David Shaffer^{1,3}, Hui-Kuan Lin^{1,2}, Zohar A. Dotan^{1,2}, Masaru Niki^{1,2}, Jason A. Koutcher³, Howard I. Scher³, Thomas Ludwig⁴, William Gerald², Carlos Cordon-Cardo² & Pier Paolo Pandolfi^{1,2}

Cellular senescence has been theorized to oppose neoplastic transformation triggered by activation of oncogenic pathways *in vitro*^{1–3}, but the relevance of senescence *in vivo* has not been established. The PTEN and p53 tumour suppressors are among the most commonly inactivated or mutated genes in human cancer including prostate cancer^{4,5}. Although they are functionally distinct, reciprocal cooperation has been proposed, as PTEN is thought to regulate p53 stability, and p53 to enhance PTEN transcription^{6–10}. Here we show that conditional inactivation of *Trp53* in the mouse prostate fails to produce a tumour phenotype, whereas complete *Pten* inactivation in the prostate triggers non-lethal invasive prostate cancer after long latency. Strikingly, combined inactivation of *Pten* and *Trp53* elicits invasive prostate cancer as early as 2 weeks after puberty and is invariably lethal by 7 months of age. Importantly, acute *Pten* inactivation induces growth arrest through the p53-dependent cellular senescence pathway both *in vitro* and *in vivo*, which can be fully rescued by combined loss of *Trp53*. Furthermore, we detected evidence of cellular senescence in specimens from early-stage human prostate cancer. Our results demonstrate the relevance of cellular senescence in restricting tumorigenesis *in vivo* and support a model for cooperative tumour suppression in which p53 is an essential failsafe protein of *Pten*-deficient tumours.

'Cellular senescence' describes a permanent form of cell cycle arrest in primary cultured cells, which can be triggered by DNA damage or activated oncogenes^{1–3}. Although it has been implicated in mediating the response to anti-tumour treatments¹¹, there is still no evidence that senescence opposes tumorigenesis.

Up to 70% of primary prostate tumours lose one *PTEN* allele and retain the other copy^{12–15}. Similarly, p53 is found completely lost or mutated almost exclusively in advanced prostate cancer^{16,17}. Because complete loss of *Pten* in the mouse seems to be crucial for the development of invasive prostate tumours^{18,19}, why human invasive prostate cancer would not also select for complete *PTEN*-loss is puzzling. Similarly, the observation that p53 is lost not at presentation but in advanced prostate cancer was surprising, because loss of p53 favours tumour initiation in many tissues. A network involving the mutual dependence of PTEN and p53 has emerged that could partly reconcile the paradox. Specifically, PTEN may protect p53 from Mdm2-mediated degradation, whereas p53 can enhance the transcription of PTEN^{6–10}. Inactivation of either gene should result in lower protein levels of the other gene. Thus, a loss of *PTEN* or p53 could mimic two genetic hits in one. To test the relevance of these observations in a well-defined model for tumour progression, we generated mice in which the *Pten* and/or *Trp53* genes are deleted specifically in the prostate. This analysis led to unforeseen

observations that contradict the above hypothesis and show the importance of cellular senescence for tumour suppression *in vivo*.

For prostate-specific inactivation, we made use of the *Cre/loxP* technique²⁰ and *Probasin-Cre4* (*Pb-Cre4*) transgenic mice expressing *Cre* after puberty (at age 7 weeks) in prostatic epithelium²¹. We obtained *Pten*^{loxP/loxP}; *Pb-Cre4* and *Trp53*^{loxP/loxP}; *Pb-Cre4* mice, hereafter referred to as *Pten*^{pc-/-} and *Trp53*^{pc-/-} (Supplementary Fig. S1). As expected, in the presence of *Pb-Cre4*, recombination of *Pten* and *Trp53* was restricted to the three prostatic lobes, namely the anterior prostate (AP), ventral prostate (VP) and dorsolateral prostate (DLP), with minor recombination occurring in seminal vesicles (Supplementary Fig. S2a).

To study early effects of *Pten* and/or *Trp53* inactivation in the prostate, mice were killed at 9 weeks of age and histopathological analysis was performed. Wild-type (WT) mice displayed normal prostate histology, whereas age-matched *Pten*^{pc-/-} littermates consistently displayed high-grade prostatic intraepithelial neoplasia (HG-PIN) in all three prostate lobes (roughly 50–60% of prostate glands affected; *n* = 10; Fig. 1a and Supplementary Fig. S2b), showing early onset of tumorigenesis immediately after *Pten* excision. In *Trp53*^{pc-/-} mice, no pathological changes (hyperplasia or PIN) were found in any prostate lobes, and their prostate glands were indistinguishable from age-matched WT mice (*n* = 10; Fig. 1b and Supplementary Fig. S2b). Up to 18 months (Supplementary Fig. S2c; not shown), there were no differences in morphology and no differences detected by immunostaining for androgen receptor (glandular epithelial marker) or p63 (basal cell marker) (Supplementary Fig. S2d).

In contrast, on combined inactivation of *Pten* and *Trp53*, we found HG-PIN in all three lobes from 100% of mice, and invasive prostate cancer in 50% of *Pten*^{pc-/-}; *Trp53*^{pc-/-} mice by age 10 weeks (*n* = 12) (Fig. 1a, b and Supplementary Fig. S2b). By 11 weeks, invasive adenocarcinoma was restricted to *Pten*^{pc-/-}; *Trp53*^{pc-/-} mutants (Fig. 1b) whereas *Pten*^{pc-/-} animals presented with invasive prostate cancer after a 4–6-month latency. Thus, in contrast to loss of *Pten*, loss of *Trp53* does not initiate prostate tumours but accelerates the progression of tumours initiated by loss of *Pten*.

We followed a cohort of 128 mice over 10 months for survival analysis, with two-weekly magnetic resonance imaging (MRI) analyses¹⁹. None of the *Pten*^{pc-/-} mutants died of prostate cancer (Fig. 1c), which is consistent with our previous study¹⁹. In contrast, reduction of *Trp53* in a *Pten*-null background markedly reduced survival in a dose-dependent manner (Fig. 1c). Notably, all *Pten*^{pc-/-}; *Trp53*^{pc-/-} mice died by 7 months of age (mean survival 5 ± 0.5 months), probably due to bladder obstruction and renal

¹Cancer Biology and Genetics Program, ²Department of Pathology, and ³Departments of Medicine, Radiology and Medical Physics, Memorial Sloan-Kettering Cancer Center, Sloan-Kettering Institute, 1275 York Avenue, New York, New York 10021, USA. ⁴Department of Anatomy and Cell Biology, Institute of Cancer Genetics, Columbia University, 1150 St Nicholas Avenue, New York, New York 10032, USA.

failure. Distant metastases were not observed in these mice or in *Pten*^{pc-/-} mice in a 2.5-year follow-up (ref. 19, and Z.C. and P.P.P., unpublished observations). By 6 months, MRI revealed masses in *Pten*^{pc-/-} prostates but not in *Trp53*^{pc-/-} or wild-type prostates (Fig. 1d). The average tumour weight in double-null mice was increased 32-fold in comparison with tumours in *Pten*^{pc-/-} mice (Fig. 1e, f) and tumours predominantly showed poorly differentiated histology (Supplementary Fig. S2c). Tumours arising in *Pten*^{pc-/-};*Trp53*^{pc-/-} mice encompassed the entire genitourinary tract and pelvis. In contrast, tumours in *Pten*^{pc-/-} mice retained clearly distinguishable urogenital organ boundaries. Thus, inactivation of *Trp53* leads to massive tumour growth and lethal prostate cancer, but only in combination with *Pten* deficiency.

We investigated the molecular basis of cooperation between *Trp53* and loss of *Pten* by using primary mouse embryonic fibroblasts (MEFs). To study the effect of acute inactivation of *Pten* and *Trp53* we infected MEFs with retroviruses expressing Cre-PURO-IRES conjugated with green fluorescent protein (GFP) (Fig. 2a and Supplementary Fig. S3a). Virus-mediated expression of Cre led to efficient recombination of loxP alleles and increased Akt activation (Supplementary Fig. S3a). *Pten* heterozygous (*Pten*^{Δ/+}) MEFs proliferated more rapidly than *Pten* WT (*Pten*^{+/+}) MEFs, but the

proliferation of *Pten*-null (*Pten*^{Δ/Δ}) MEFs was similar to that of the WT (Fig. 2a). *Pten*-null MEFs exhibited the distinctive morphology of senescent cells (flattened large cells; not shown) and were positive for senescence-associated β-galactosidase (SA-β-Gal), a hallmark of senescent cells²² (Fig. 2a and Supplementary Fig. S5a). To determine whether senescence was related to the selection process or to retroviral integration, we made use of transient episomal delivery by means of adenovirus-Cre (Ad-Cre) infection. We found that the loss of one *Pten* allele resulted in accelerated growth compared with that of wild-type cells, yet the loss of both *Pten* alleles led to lower proliferation and high β-Gal staining (Supplementary Figs S3b, c and S5d). To ensure that these findings were unrelated to Cre activity, we attempted *Pten* knockdown by means of RNA interference (RNAi). A decrease in *Pten* levels by 50% in heterozygous MEFs did indeed result in lowered proliferation and senescence induction, confirming not only the Cre-independence of senescence, but also suggesting that acute loss of *Pten* below the heterozygosity threshold can elicit a cellular senescence response (Fig. 2b, and Supplementary Figs S3d and S5b).

Next we assessed the status of the p53 senescence pathway. Acute loss of *Pten* resulted in Akt activation and consistent induction of p53 levels (Fig. 2c) but not phosphorylation of p53 serine 15 (Sup-

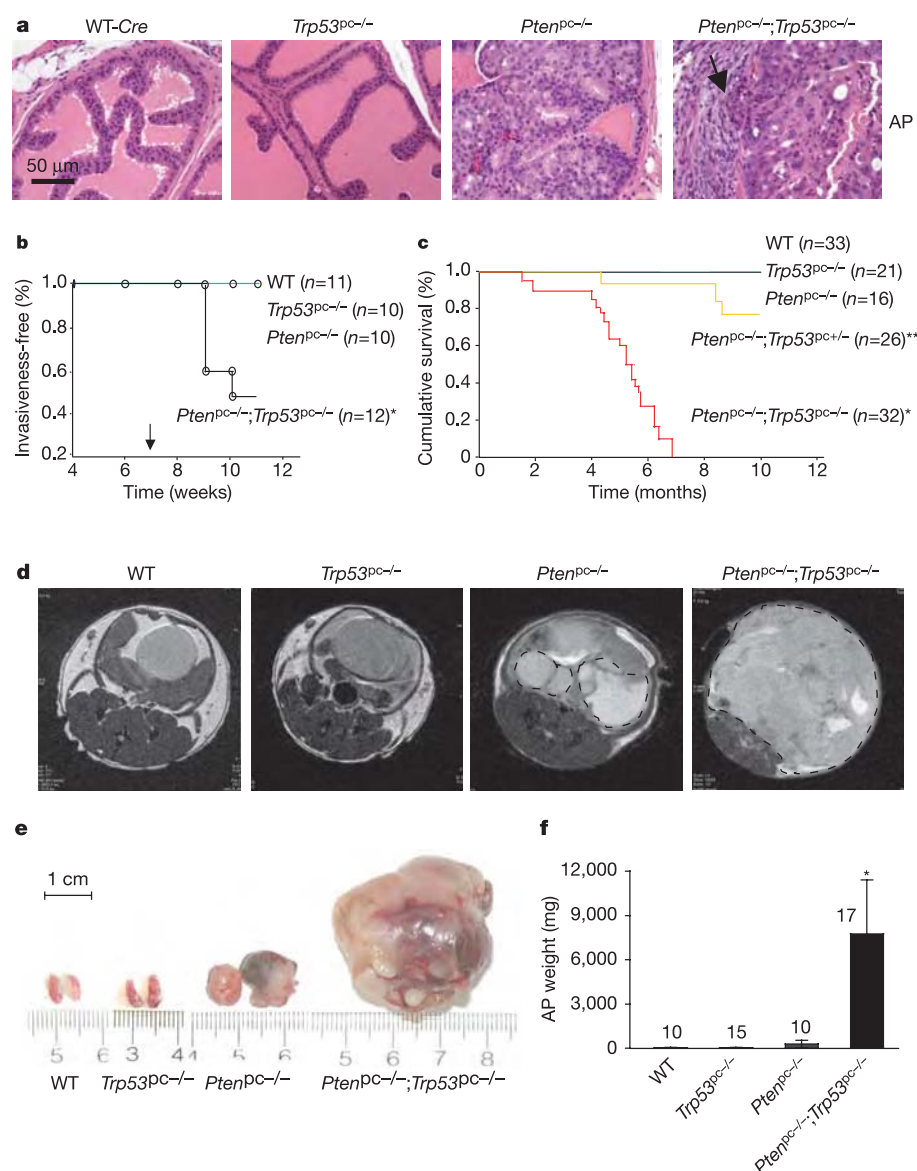


Figure 1 | Loss of *Trp53* does not initiate prostate tumours but renders *Pten*-deficient carcinomas lethal. **a**, Histopathological analysis (haematoxylin/eosin staining) of anterior prostates (AP) in WT, *Pten* and *Trp53* single and double mutants at 9 weeks of age reveals normal glands in WT and *Trp53*^{pc-/-} mice but PIN lesions in *Pten*^{pc-/-} mice and invasion (arrow) in *Pten*^{pc-/-};*Trp53*^{pc-/-} mice. **b**, Disease-free survival curve (Kaplan–Meier plot) for prostate cancer. Adenocarcinoma was found only in the *Pten*^{pc-/-};*Trp53*^{pc-/-} cohort ($P < 0.05$). The arrow indicates puberty. **c**, Cumulative survival analysis. A statistically significant decrease in lifespan ($P < 0.0001$) compared with the *Pten*^{pc-/-} cohort was found for the *Pten*^{pc-/-};*Trp53*^{pc-/-} cohort (asterisk) and for the *Pten*^{pc-/-};*Trp53*^{pc-/-} cohort (double asterisk). **d–f**, MRI of AP tumours (dashed circles) at 23–31 weeks (**d**) and their actual sizes (**e**) and weights (**f**) after biopsy. Error bars in **f** indicate s.d. of AP weight for the numbers of mice indicated above the bars.

plementary Fig. S3f). Upregulation of p21 and the senescence marker plasminogen activator inhibitor-1 (PAI-1) was observed in *Pten*-deficient MEFs but not in *Pten/Trp53* double-null MEFs (Fig. 2c and Supplementary Fig. S3g). Furthermore, loss of *Trp53* led to a marked increase in proliferation in the *Pten*-null background (Fig. 2d and Supplementary Fig. S3e). In addition, SA- β -Gal activity in *Pten* ^{Δ/Δ} MEFs decreased sharply in a p53 dose-dependent manner and was undetectable in double-null MEFs. These findings confirm that escape from senescence is essential for *Pten* ^{Δ/Δ} MEFs to realize full proliferative potential (Fig. 2d). Staining by TdT-mediated dUTP nick end labelling (TUNEL) revealed no difference in apoptosis between *Pten*-null and *Pten/Trp53* double-null MEFs (2 and 4 days, not shown). Finally, we tested whether p53 antagonizes cellular transformation of MEFs²³. Complete inactivation of *Pten* and *Trp53* resulted in transformation of MEFs, whereas transformation potential decreased in a *Pten* or *Trp53* copy-dependent manner (Fig. 2e). These findings suggested that a p53-mediated senescence response restricts cell growth after loss of *Pten*.

Next we investigated whether the stability of p53 was altered on

loss of *Pten* as reported previously⁹. In fact, we observed a prolonged rather than a reduced half-life of p53 protein in *Pten*-null compared with WT MEFs (Fig. 3a). This could be explained by the fact that complete loss of *Pten* resulted in increased accumulation of p19^{Arf} (Fig. 3b), which is implicated in both the induction of cellular senescence and p53 stabilization through Mdm2 inhibition^{1,2}. p16^{INK4a} protein levels were also increased (Fig. 3b). In particular we found that the half-life of p21 was prolonged (Fig. 3a), indicating that p53-independent (post-transcriptional) mechanisms might also affect the accumulation of p21. Nevertheless, p53 was essential for p21 induction, because p21 was almost undetectable in *p53/Pten* double-null cells (Fig. 2c). Moreover, as reported previously²⁴, an activated form of Akt (myristoylated Akt) elicited growth arrest and cellular senescence in primary MEFs, supporting the notion that senescence caused by acute loss of *Pten* is at least partly due to Akt activation (Fig. 3c and Supplementary Fig. S5c).

To determine whether these observations were relevant *in vivo* to prostate cancer we performed immunohistochemical analysis (IHC). Staining of *Pten* revealed that *Pten* was efficiently and specifically

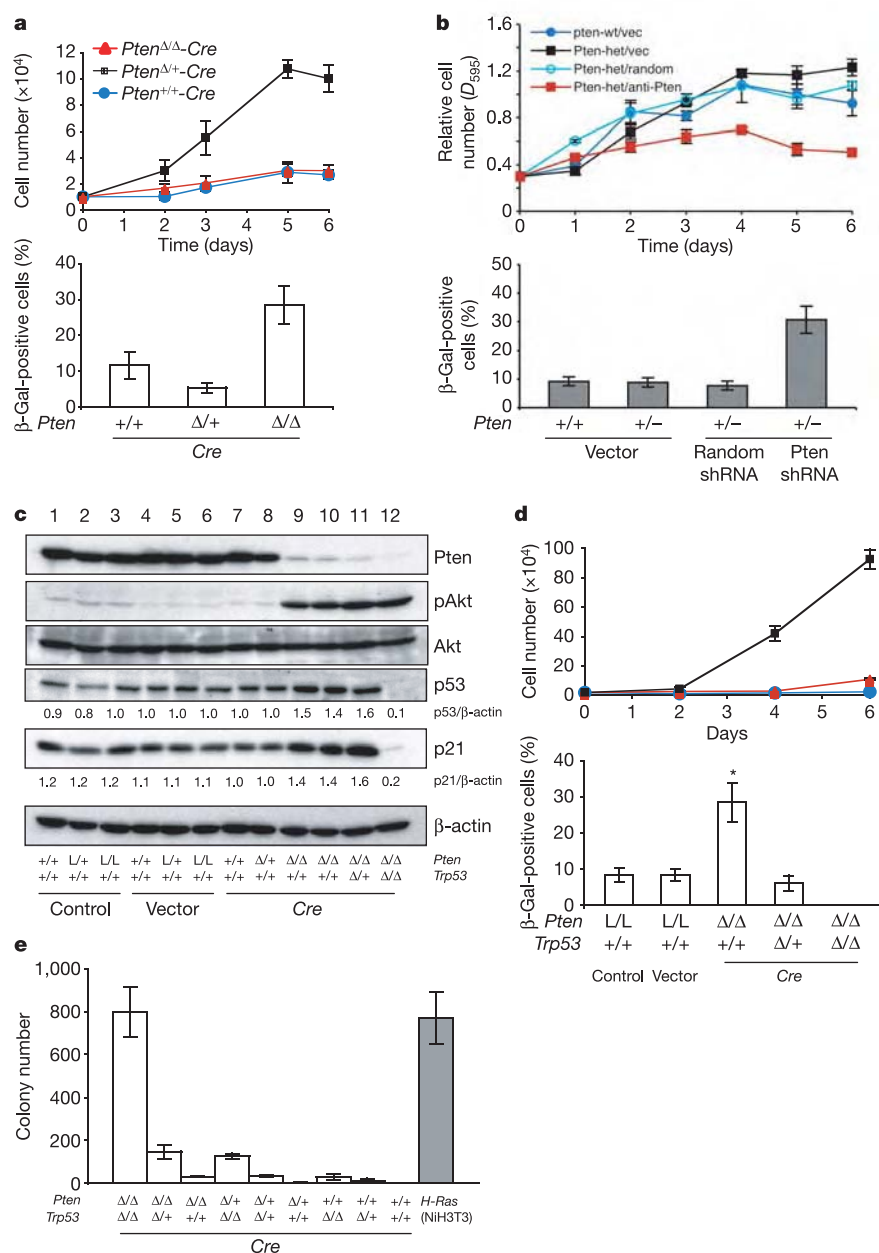


Figure 2 | Acute loss of *Pten* triggers the p53-dependent senescence pathway in primary mouse embryonic fibroblasts (MEFs). **a**, Top: growth curves of primary MEFs, infected with retroviral Cre (with selection) and followed over a 6-day period. Bottom: cellular senescence assay of cells from **a**. **b**, Top: growth curves of primary MEFs infected with *Pten* shRNA (with selection) and followed over a 6-day period. Bottom: cellular senescence assay of cells from **b**. **c**, Western blots of MEF lysates. Numbers indicate densitometrically determined protein levels relative to β -actin. **d**, Growth curves (top) and senescence staining (bottom) of *Pten* ^{Δ/Δ} ; *Trp53* ^{Δ/Δ} double-null MEFs (black squares), *Pten* ^{Δ/Δ} ; *Trp53* ^{$\Delta/+$} MEFs (red triangles) and *Pten* ^{Δ/Δ} MEFs (blue circles). Error bars indicate s.d. L/L indicates *Pten*^{*loxP/loxP*}. **e**, Transformation assay of all combinations of *Pten* and *Trp53* inactivation. H-Ras-infected NIH 3T3 cells served as positive control (grey bar). Error bars indicate s.d. for a representative experiment performed in triplicate.

deleted and unaffected by *Trp53* status (Supplementary Fig. S4a). Akt activation and membrane recruitment were correlated only with loss of *Pten* in *Pten*^{pc-/-} and *Pten*^{pc-/-}; *Trp53*^{pc-/-} mice, and not with p53 status (Supplementary Fig. S4a). In parallel, frozen prostate sections were stained for β -Gal activity. Whereas *Pten*-null prostates contained large numbers of senescent cells ('blue glands'), WT prostates had very few (Fig. 4a and Supplementary Fig. S6b). Quantification revealed a 20-fold increase in senescent epithelial cells at 11 weeks compared with WT (Fig. 4b), but senescent cells were markedly decreased in the double-null prostates (Fig. 4a, b and Supplementary Fig. S6b). In addition, no cellular senescence was detected after heterozygous inactivation of *Pten* in *Pten*^{pc+/-} single or compound mutants (Supplementary Fig. S6a), which is consistent with our MEF results (Supplementary Fig. S5a, b, d) and in full agreement with the fact that partial or complete inactivation of *Trp53* did not accelerate the PIN phenotype of *Pten*^{pc+/-} mutants (not shown). Apoptosis in prostates from various mutants remained comparable (Fig. 4c), whereas senescence of epithelial cells was accompanied by inverse changes in cellular proliferation (Fig. 4d); this was confirmed by double labelling of *Pten*-mutant and double mutant glands with Ki-67 and β -Gal (Supplementary Fig. S4b).

We next studied whether p53 is upregulated on *Pten* inactivation *in vivo*. Indeed, we found increased accumulation of p53 in the nuclei of prostate epithelial cells from *Pten*^{pc-/-} mice but not from WT (or *Pten*^{pc-/-}; *Trp53*^{pc-/-} mice (Fig. 4a). Lysates from *Pten*^{pc-/-} prostates showed more than tenfold p53 induction by immunoblotting (Fig. 4e). Importantly, p19^{Arf} and p21 protein levels were also

increased in *Pten*^{pc-/-} mice (Fig. 4a), confirming *in vivo* activation of the p19^{Arf} → p53 → p21 senescence pathway.

To address the relevance of these findings to human cancer we stained for senescence in specimens from early-stage human prostate cancer ($n = 12$; Supplementary Table S1). Strong β -Gal staining was observed specifically in regions of prostate hyperplasia/PIN in some glands (3 of 12 samples, magnification $\times 100$), but never in areas of frank tumour. Weak staining was observed in a further four samples at higher magnification ($\times 400$), predominantly in areas of prostate hyperplasia/PIN and rarely in areas of carcinoma (Fig. 4f; Supplementary Table S1).

Prostate cancer, the second leading cause of cancer-related deaths in males (after lung cancer), strikes one in six men in the USA^{25,26} and involves alterations of multiple tumour suppressor genes²⁷. Our data provide insight into the molecular relationship between *Pten* and p53 in the prostate by demonstrating that loss of *Pten* induces p53 function (Fig. 4g). Because it has remained unclear whether loss of p53 function is associated with the initiation or progression of prostate cancer, our data firmly associate loss of *Trp53* with a crucial role in prostate cancer progression (Fig. 4h). Our results show that acute loss of *Pten* does not result in decreased, but increased, p53 levels and function. Conversely, loss of *Trp53* has no effect on *Pten* expression *in vitro* and in prostatic epithelium, whereas in combination with loss of *Pten* a lethal acceleration of prostate cancer is observed. Thus, regarding the *Pten*-p53 network, our data lend no support to a 'two-in-one' hit model of tumorigenesis, but as both genes need to be ablated for maximal disease progression they suggest a 'one-by-one' hit model for the genetic interaction of these major tumour suppressor genes. Our findings have therapeutic implications, because they indicate that *PTEN*-deficient prostate cancer might benefit from drugs that potentiate p53 activation in favour of a cellular senescence programme^{28,29} (Fig. 4g). The fact that acute homozygous (not heterozygous) loss of *Pten* triggers a p53-dependent cellular senescence programme *in vivo* provides a plausible explanation for why human prostate cancer at presentation does not select for complete loss of *PTEN* and thus highlights the relevance of *PTEN* haploinsufficiency for prostate cancer initiation. Furthermore, the finding that complete p53 inactivation alone has no phenotypic consequences in prostate could explain why loss of p53 is preferentially selected for in advanced prostate cancer where in addition it could allow the tumour to reach maximal proliferation through *PTEN* loss of heterozygosity.

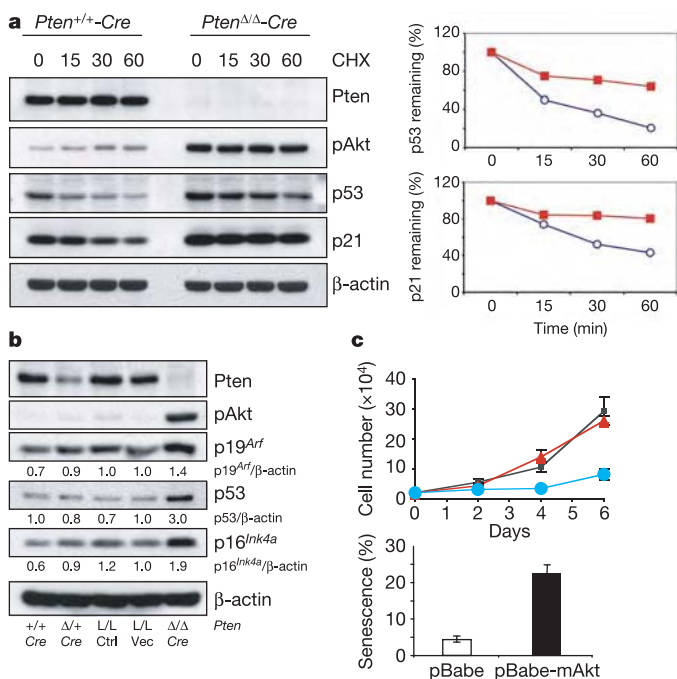


Figure 3 | Acute loss of *Pten* results in ARF upregulation and p53/p21 stabilization in primary MEFs. **a**, Left: inhibition of protein synthesis by cycloheximide (CHX) and western blotting at the indicated times (minutes) shows that the half-life of p53 and p21 proteins is prolonged on acute loss of *Pten* in MEFs. Right: quantification of p53 and p21 half-life from the western blots, normalized to β -actin. Blue circles, *Pten*^{+/+-Cre}; red squares, *Pten*^{ΔΔ-Cre}. **b**, Western blotting demonstrates upregulation of p19^{Arf} protein after acute *Pten* loss in MEFs. *Pten*^{+/+-Cre}, *Pten*^{ΔΔ-Cre}, *Pten*^{loxP/loxP} (L/L, Ctrl) and *Pten*^{loxP/loxP}-MSCV vector (*Pten*^{L/L}-vec) serve as controls. **c**, Expression of myristoylated Akt (mAkt) in MEFs induces growth arrest (top) and cellular senescence (bottom). Black squares, control; red triangles, pBabe; blue circles, pBabe-mAkt. Error bars represent s.d. for a representative experiment performed in triplicate.

METHODS

***Pten* and *Trp53* mutant mice.** *Pten*^{loxP/loxP} mice were generated as described previously¹⁹, and *Trp53*^{loxP/loxP} mice were generated with a similar strategy (Supplementary Fig. S1; details are available from the authors). Female *Pten*^{loxP/loxP}; *Trp53*^{loxP/loxP} mice were crossed with male *PC-Cre* transgenic mice²¹ for the prostate-specific deletion of *Pten* and *Trp53*. For genotyping, tail DNA was subjected to polymerase chain reaction analysis with the following primers. For *Pten*^{loxP/loxP}, primer 1 (5'-AAAAGTCCCTGCTGATGATTGTG-3') and primer 2 (5'-TGTTTTGACCAATTAAGTAGGCTGTG-3') were used. To detect the deleted allele, *Lptn3* primer (5'-TTCTCTTGAGCACTGTTTACAGGC-3') and primer 1 were used. For *Trp53*^{loxP/loxP}, primer 1 (5'-GAGACGCTGAGTCCGGTTCCTCC-3') and primer 2 (5'-GCAAGA GGTACTTTGGGTGAAGTC-3') were used.

Constructs. pMSCV-Cre-PIG was constructed by subcloning *Cre* into pMSCV-PURO-IRES-GFP (MSCV-PIG) vector (a gift from S. W. Lowe). The Nuclear Localization Sequence fusion Cre recombinase (NLS-Cre) was excised from pTZ-CreN vector (a gift from L. Nitschke).

Cell proliferation, transformation, senescence and apoptosis assays. Primary MEFs were prepared as described previously from individual embryos of various genotypes³⁰, and infected with retroviruses expressing Cre-PURO-IRES-GFP, myristoylated Akt (mAkt) or corresponding control viruses. To prepare retroviral particles, 2×10^6 Phoenix cells were plated per 10-cm culture dish, and then pMSCV-Cre-PURO-IRES-GFP or empty vector was transfected with Lipofectamine 2000 (Invitrogen). For infection, MEFs at passage 2 were plated at a density of $(3-4) \times 10^5$ cells per 10-cm culture dish and infected by virus from Phoenix cells 48 h after transfection. After selection in the presence of $3 \mu\text{g ml}^{-1}$

puromycin (Sigma), MEFs at passage 6 were used for growth curves, western blotting and senescence analysis. To determine senescence, MEFs were plated at 10^4 cells per well of a six-well plate in triplicate, and after 4 days SA- β -Gal activity was revealed with the senescence detection kit (Calbiochem) and quantified

(more than 200 cells per sample). For prostate tissue, frozen sections $6\ \mu\text{m}$ thick were stained for β -Gal as above. For transformation assay, selected MEFs (3×10^4) at passage 6 were suspended in medium containing 0.3% agar onto solidified 0.6% agar per well of a six-well plate, and the number of colonies was

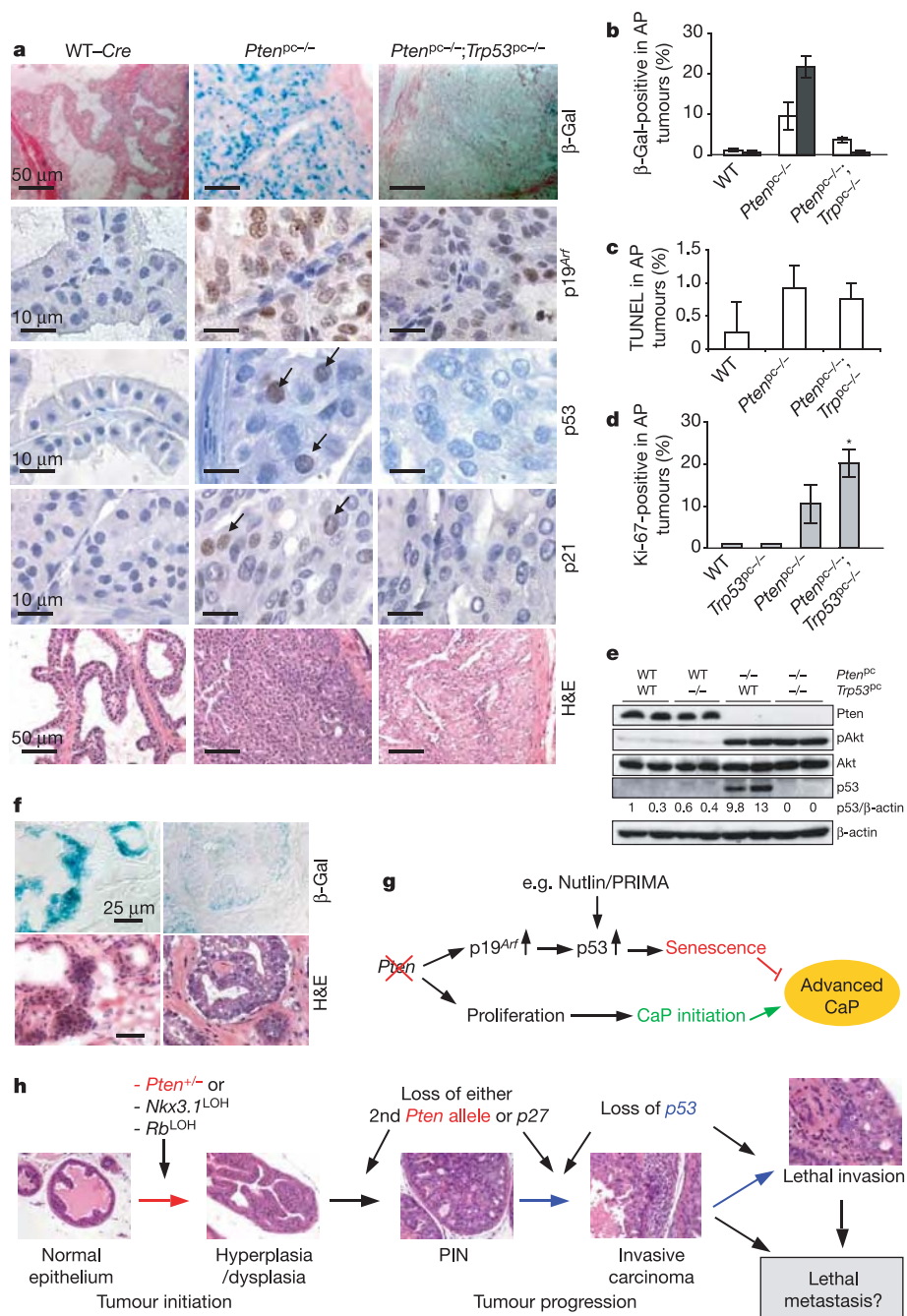


Figure 4 | The p53-dependent cellular senescence pathway restricts tumorigenesis in *Pten*-deficient prostates. **a**, Senescence and histopathological analysis of 11-week-old prostates, stained as indicated. H&E, haematoxylin/eosin. Scale bars, 50 μm (β -Gal and haematoxylin/eosin stain) and 10 μm (p53, p19^{Arf} and p21). **b**, Quantification of the β -Gal staining seen on AP sections at 8 weeks (open bars) and 11 weeks (filled bars). Representative sections from three mice were counted for each genotype. **c**, Quantification of TUNEL assay for apoptosis in the AP at 11 weeks (of more than two mice per genotype). **d**, Quantification of Ki-67 staining of 11-week-old AP done as in **b**. Error bars in **b–d** represent s.d. for a representative experiment performed in triplicate. Asterisk indicates statistical significance between *Pten*^{pc-/-} and *Pten*^{pc-/-}; *Trp53*^{pc-/-} double mutants ($P < 0.05$). **e**, Western blot analysis of AP tissue from each genotype at 11 weeks shows Akt activation and p53 induction in *Pten*^{pc-/-}

mice. **f**, Senescence (β -Gal) and histopathological (haematoxylin/eosin stain) analysis of cryosections from human radical prostatectomy samples, showing strong (+++) β -Gal staining from a hyperproliferative (preneoplastic) prostate gland (left) or weak (+) β -Gal staining from a neoplastic gland from a different patient (right). Scale bars, 25 μm . **g**, A model for *Pten*-deficient tumorigenesis, p53 cooperativity and its therapeutic implications. *Pten*-deficient tumours may benefit from treatment with drugs that restore the activity of mutated p53 (for example, PRIMA²⁸) or that stabilize WT p53 through the inactivation of MDM2 (for example, Nutlin²⁹). **h**, A model for prostate tumour initiation, development and progression synergistically contributed by *Pten*, *Trp53* and other genes. Loss of *Trp53* accelerates cancer progression by a senescence escape mechanism in *Pten*-deficient tumours.

assessed after 21 days. For apoptosis analysis, dewaxed and rehydrated paraffin sections were treated for with the *in situ* Cell Death Detection Kit (Roche). Apoptotic cells were identified by positive TUNEL staining. A total of 500 cells were counted from five different fields for each section, and the analysis was repeated at least twice for each genotype.

Short hairpin RNA (shRNA). Pten-directed shRNA (5'-GATCCCCAGACCA TAACCCACCAAGTCAAGAGACTGTGGTGGGTATGGTCTTTT-3') was designed using the Oligoengine RNAi design tool; the resulting oligonucleotides were subcloned into the pSUPER-puro vector (Oligoengine) and transfected into Phoenix packaging cells. The random oligonucleotide (5'-GATCCCCAAGGAGACGAGCAAGAGAATTCAAGAGATTCTTGTCTGTC TCCTTTT-3') was used as control. Primary *Pten* WT or heterozygous MEFs at passage 2 were infected, selected (3 days in 2 $\mu\text{g ml}^{-1}$ puromycin) and plated at 2.5×10^4 cells per well in 12-well dishes in triplicates for determination of growth curves by spectroscopic measurement of crystal violet uptake. β -Gal activity was measured on day 4 of the growth curve with cells treated identically and in parallel.

Western blot and immunohistochemistry. MEF lysates were prepared with RIPA buffer (1 $\times$ PBS, 1% Nonidet P40, 0.5% sodium deoxycholate, 0.1% SDS and protease inhibitor cocktail (Roche)). The following antibodies were used for western blotting: mouse monoclonal anti-Pten (6H2.1; Cascade BioScience), rabbit polyclonal anti-p19^{Arf} (Ab-1; Calbiochem), rabbit polyclonal anti-p53 (CM5; Novocastra), rabbit polyclonal anti-Akt and anti-phospho-serine 473 of Akt (Cell Signalling), rabbit polyclonal anti-p21 (C-19; Santa Cruz), rabbit polyclonal anti-p16 (M-156; Santa Cruz), rabbit polyclonal anti-PAI-1 (H-135; Santa Cruz) and mouse monoclonal anti- β -actin (AC-74; Sigma). To determine p53/p21 protein half-life, MEFs (*Pten*^{+/+}-Cre and *Pten* ^{$\Delta\Delta$} -Cre, at 80% confluence and equal passage number) were treated with 30 $\mu\text{g ml}^{-1}$ cycloheximide (Sigma) and harvested at the indicated times for western blot analysis. For prostate analysis, protein extracts were prepared by grinding prostate tissues in RIPA buffer at a ratio of 1 mg per 5 μl . After brief sonication and centrifugation, the supernatant was collected for western blotting. Antibodies used were rabbit polyclonal anti-PTEN (9552; Cell Signalling). Antibodies against p53, pAkt and Akt were as above. For immunohistochemistry (IHC), tissues were fixed in 10% formalin and embedded in paraffin in accordance with standard procedures. Sections were stained for phospho-Akt (Ser 473) antibody (Cell Signalling), PTEN (Ab-2; NeoMarkers), Ki-67 (Novocastra), p19^{Arf} (ab80-50; Abcam), p21 (F-5; Santa Cruz), p53 (FL-393; Santa Cruz), androgen receptor (N-20; Santa Cruz) and p63 (550025; Becton Dickinson Transduction Lab).

MRI. Individual mice were subjected to MRI assessment for the detection of prostate tumours. In brief, mice were anaesthetized with 2% isoflurane, and images were obtained on a Bruker 4.7-T 40-cm bore magnet with a commercial 7-cm inner diameter birdcage coil similar to the protocol described previously¹⁹. Low-resolution sagittal and axial scout images were obtained initially, followed by high-spatial-resolution T_2 -weighted axial images (repetition interval (T_R) = 3,800 ms, effective echo time (T_E) = 35 ms, eight echoes per phase encoding step, spatial resolution = 1.0 mm slice thickness $\times$ 112 μm $\times$ 112 μm in plane resolution, and four repetitions of data acquisition for 8–9 min of imaging time). Invasiveness-free and cumulative survival curves were obtained with Kaplan–Meier analysis as described previously¹⁹.

Received 23 February; accepted 15 June 2005.

1. Lowe, S. W., Cepero, E. & Evan, G. Intrinsic tumour suppression. *Nature* **432**, 307–315 (2004).
2. Serrano, M. & Blasco, M. A. Putting the stress on senescence. *Curr. Opin. Cell Biol.* **13**, 748–753 (2001).
3. Campisi, J. Cellular senescence as a tumour-suppressor mechanism. *Trends Cell Biol.* **11**, S27–S31 (2001).
4. Vogelstein, B., Lane, D. & Levine, A. J. Surfing the p53 network. *Nature* **408**, 307–310 (2000).
5. Di Cristofano, A. & Pandolfi, P. P. The multiple roles of PTEN in tumour suppression. *Cell* **100**, 387–390 (2000).
6. Stambolic, V. et al. Regulation of PTEN transcription by p53. *Mol. Cell* **8**, 317–325 (2001).
7. Mayo, L. D. & Donner, D. B. A phosphatidylinositol 3-kinase/Akt pathway

promotes translocation of Mdm2 from the cytoplasm to the nucleus. *Proc. Natl Acad. Sci. USA* **98**, 11598–11603 (2001).

8. Zhou, B. P. et al. HER-2/neu induces p53 ubiquitination via Akt-mediated MDM2 phosphorylation. *Nature Cell Biol.* **3**, 973–982 (2001).
9. Freeman, D. J. et al. PTEN tumour suppressor regulates p53 protein levels and activity through phosphatase-dependent and -independent mechanisms. *Cancer Cell* **3**, 117–130 (2003).
10. Trotman, L. C. & Pandolfi, P. P. PTEN and p53: who will get the upper hand? *Cancer Cell* **3**, 97–99 (2003).
11. Schmitt, C. A. et al. A senescence program controlled by p53 and p16INK4a contributes to the outcome of cancer therapy. *Cell* **109**, 335–346 (2002).
12. Suzuki, H. et al. Interfocal heterogeneity of PTEN/MMAC1 gene alterations in multiple metastatic prostate cancer tissues. *Cancer Res.* **58**, 204–209 (1998).
13. Feilother, H. E., Nagai, M. A., Boag, A. H., Eng, C. & Mulligan, L. M. Analysis of PTEN and the 10q23 region in primary prostate carcinomas. *Oncogene* **16**, 1743–1748 (1998).
14. Muller, M., Rink, K., Krause, H. & Miller, K. PTEN/MMAC1 mutations in prostate cancer. *Prostate Cancer Prostatic Dis.* **3**, S32 (2000).
15. Hermans, K. G. et al. Loss of a small region around the PTEN locus is a major chromosome 10 alteration in prostate cancer xenografts and cell lines. *Genes Chromosom. Cancer* **39**, 171–184 (2004).
16. Qian, J. et al. Loss of p53 and c-myc overrepresentation in stage T(2–3)N(1–3)M(0) prostate cancer are potential markers for cancer progression. *Mod. Pathol.* **15**, 35–44 (2002).
17. Navone, N. M. et al. p53 mutations in prostate cancer bone metastases suggest that selected p53 mutants in the primary site define foci with metastatic potential. *J. Urol.* **161**, 304–308 (1999).
18. Di Cristofano, A., De Acetis, M., Koff, A., Cordon-Cardo, C. & Pandolfi, P. P. Pten and p27^{KIP1} cooperate in prostate cancer tumour suppression in the mouse. *Nature Genet.* **27**, 222–224 (2001).
19. Trotman, L. C. et al. Pten dose dictates cancer progression in the prostate. *PLoS Biol.* **1**, 385–396 (2003).
20. Jonkers, J. & Berns, A. Conditional mouse models of sporadic cancer. *Nature Rev. Cancer* **2**, 251–265 (2002).
21. Wu, X. et al. Generation of a prostate epithelial cell-specific Cre transgenic mouse model for tissue-specific gene ablation. *Mech. Dev.* **101**, 61–69 (2001).
22. Dimri, G. P. et al. A biomarker that identifies senescent human cells in culture and in aging skin *in vivo*. *Proc. Natl Acad. Sci. USA* **92**, 9363–9367 (1995).
23. Weinberg, R. A. The cat and mouse games that genes, viruses, and cells play. *Cell* **88**, 573–575 (1997).
24. Miyauchi, H. et al. Akt negatively regulates the *in vitro* lifespan of human endothelial cells via a p53/p21-dependent pathway. *EMBO J.* **23**, 212–220 (2004).
25. Stewart, S. L., King, J. B., Thompson, T. D., Friedman, C. & Wingo, P. A. Cancer mortality surveillance—United States, 1990–2000. *MMWR Surveill. Summ.* **53**, 1–108 (2004).
26. Levi, F., Lucchini, F., Negri, E., Boyle, P. & La Vecchia, C. Leveling of prostate cancer mortality in Western Europe. *Prostate* **60**, 46–52 (2004).
27. Abate-Shen, C. & Shen, M. M. Molecular genetics of prostate cancer. *Genes Dev.* **14**, 2410–2434 (2000).
28. Bykov, V. J. et al. Restoration of the tumour suppressor function to mutant p53 by a low-molecular-weight compound. *Nature Med.* **8**, 282–288 (2002).
29. Vassilev, L. T. et al. *In vivo* activation of the p53 pathway by small-molecule antagonists of MDM2. *Science* **303**, 844–848 (2004).
30. Maeda, T. et al. Role of the proto-oncogene *Pokemon* in cellular transformation and ARF repression. *Nature* **433**, 278–285 (2005).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank T. Maeda and T. Jacks for helpful suggestions; D. Peeper, C. Schmitt and M. Serrano for exchanging unpublished data and coordinating the submission of manuscripts; N. Hay, U. Greber and S. Hemmi for reagents; L. Cai and L. DiSantis for critical reading and editing of the manuscript; other members of the Pandolfi lab for advice and discussion; K. Manova and C. Farrell from the Molecular Cytology Core Facility for assistance with IHC analysis; and C. Le, C. Matei, D. Prociassi and I. Buchanan for MRI analysis. This work was supported by NIH grants to P.P.P.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.P.P. (p-pandolfi@ski.mskcc.org).

A cytokinesis furrow is positioned by two consecutive signals

Henrik Bringmann¹ & Anthony A Hyman¹

The position of the cytokinesis furrow in a cell determines the relative sizes of its two daughter cells as well as the distribution of their contents. In animal cells, the position of the cytokinesis furrow is specified by the position of the mitotic spindle¹. The cytokinesis furrow bisects the spindle midway between the microtubule asters, at the site of the microtubule-based midzone, producing two daughter cells. Experiments in some cell types have suggested that the midzone positions the furrow^{2,3}, but experiments in other cells have suggested that the asters position the furrow^{4,5}. One possibility is that different organisms and cell types use different mechanisms to position the cytokinesis furrow. An alternative possibility is that both asters and the midzone contribute to furrow positioning^{6,7}. Recent work in *C. elegans* has suggested that centrosome separation and the midzone are implicated in cytokinesis⁸. Here we examine the relative contributions of different parts of the mitotic spindle to positioning of the cytokinesis furrow in the *C. elegans* zygote. By spatially separating the spindle midzone from one of the asters using an ultraviolet laser, we show that the cytokinesis furrow is first positioned by a signal determined by microtubule asters, and then by a second signal that is derived from the spindle midzone. Thus, the position of the cytokinesis furrow is specified by two consecutive furrowing activities.

A mitotic spindle contains two structures implicated in cytokinesis furrow positioning: asters, which are radial arrays of centrosome-nucleated microtubules, and the midzone, which forms between the separating chromatin at anaphase and consists of non-kinetochore spindle microtubules. A mitotic spindle is an inherently symmetric structure. A furrow-positioning cue from the asters would position the furrow midway between them. A midzone cue would position a furrow at the same place. In order to separate the contributions of the midzone and the asters to cytokinesis, the two structures must be spatially separated. In one-cell *C. elegans* embryos, the spindle forms in the middle of the cell at metaphase. At anaphase, cortical pulling forces pull on the microtubule asters, separating the two spindle poles⁹.

We took advantage of these pulling forces to design an experiment in which the position midway between the two asters was different from the position of the spindle midzone. An embryo was observed until the onset of anaphase (the time at which the midzone forms¹⁰). One aster was then separated from its associated chromatin using an ultraviolet laser, creating a cell with one isolated aster and the other aster still attached to the midzone. Cortical pulling forces moved the asters to opposite poles of the cell, positioning the spindle 'midzone' roughly one-third of the way along the aster-to-aster-distance (Fig. 1a–d and Supplementary Videos 1, 2). Thus, the position of the midzone was different from the position midway between the two asters. We call this procedure asymmetric spindle severing (ASS). Following ASS, cytokinesis furrow ingression started midway between the asters. However, the furrow did not complete midway

between the asters: a second furrow formed at the cell cortex closest to the midzone (Fig. 1e–g and Supplementary Videos 3–5). The two distinct furrows then met and cytokinesis completed. Thus, two furrows are observed after ASS, one between the asters and one directed towards the midzone. Both furrows contribute to the final position of the cleavage plane.

Using cylindrical cells, it has been shown that the mitotic apparatus can induce multiple furrows if it is successively displaced along the long axis of the cell¹¹. A furrow can even be produced after removing the mitotic apparatus by aspiration before the onset of furrowing¹². It could be that the spindle midzone specifies both furrows, one before and one after its displacement. Alternatively, the asters might specify the first furrow. To resolve this problem, we compared the effects of separating either the anterior or the posterior aster. The isolated aster moves further towards the poles of the cell compared with the aster that is still attached to the midzone. Thus, the position midway between the asters is different after anterior and posterior ASS. This difference is reflected in the difference of the position of the first furrow (see Fig. 2, the position of the first furrow is $53.7 \pm 0.9\%$ along the embryo length after anterior ASS and $56.5 \pm 0.6\%$ along the embryo length after posterior ASS; mean $\pm$ s.e.m.; $P = 0.041$). After ASS, the position of the first furrow is thus dependent on the position of the asters.

We also shifted the first furrow using monopolar spindles, as it has been shown that either an isolated aster or a monopolar spindle is sufficient to specify a cleavage plane^{4,13}. To generate monopolar spindles, we repeated ASS and subsequently disintegrated the separated aster using the ultraviolet laser (Fig. 2 and Supplementary Videos 6, 7). Under these conditions, the first furrow was established further away from the remaining aster (ASS with intact aster, $12.8 \pm 0.2 \mu\text{m}$; ASS followed by aster disintegration, $18.0 \pm 0.7 \mu\text{m}$), at a position that was not related to the midzone position before ASS. The cleavage plane then corrected towards the midzone. The spatial shift of the first furrow away from the intact aster demonstrates that the furrow position is not defined before spindle severing, and is thus determined by microtubule asters. Taken together, these experiments show that the cytokinesis cleavage plane is specified by two consecutive signals derived from distinct components of the mitotic spindle: the microtubule asters and the spindle midzone.

Cleavage furrow formation is driven by actomyosin contraction, but little is known about the molecular mechanisms specifying where and when contraction will occur. To learn about the molecular control of redundant cleavage plane positioning, we assayed genes with known and potential roles in cytokinesis for their effects on the two furrows. Genes leading to defects in cytokinesis were chosen from the literature and genome-wide screens (see Supplementary Table 1). The gene products were depleted either by RNA interference (RNAi) or using established mutants, subjected to the ASS procedure, and analysed for the specification of the aster-positioned

¹Max Planck Institute for Molecular Cell Biology and Genetics, Pfotenhauer Strasse 108, 01307 Dresden, Germany.

furrow (the first furrow), the specification of the midzone-positioned furrow (the second furrow), the ability to correct the aster-positioned furrow, and the extent of ingression of both furrows.

Here we describe four categories of cleavage furrow defects (Fig. 3 and Supplementary Videos 8–13). Supplementary Table 1 contains details of the genes tested. The first class consisted of *rho-1*, *nmy-2*, *cyk-1*, *mlc-4*, *let-21*, *pfn-1* and *act-4*. In this class, neither furrow formed. Genes in this class were required for general aspects of actomyosin activity, and have previously been shown to prevent

furrow formation^{14–16}. We did not further analyse these genes. The second class consisted of *air-2*, *zen-4* and *cyk-4* (Fig. 3b). These genes have known roles in spindle midzone formation and completion of cytokinesis^{17–19}. After ASS the aster-positioned furrow ingressed weakly compared with wild-type embryos. A midzone-positioned furrow did not form. The aster-positioned furrow regressed and cytokinesis failed. The weak ingression of the aster-positioned furrow indicates that these genes have general roles in furrow ingression.

The third class of genes leading to cleavage furrow defects consisted of *spd-1* and *kpl-7* (Fig. 3c). Both of these genes are known to be required for midzone formation, but cytokinesis completes in one-cell embryos^{9,20}. After ASS, embryos in this class

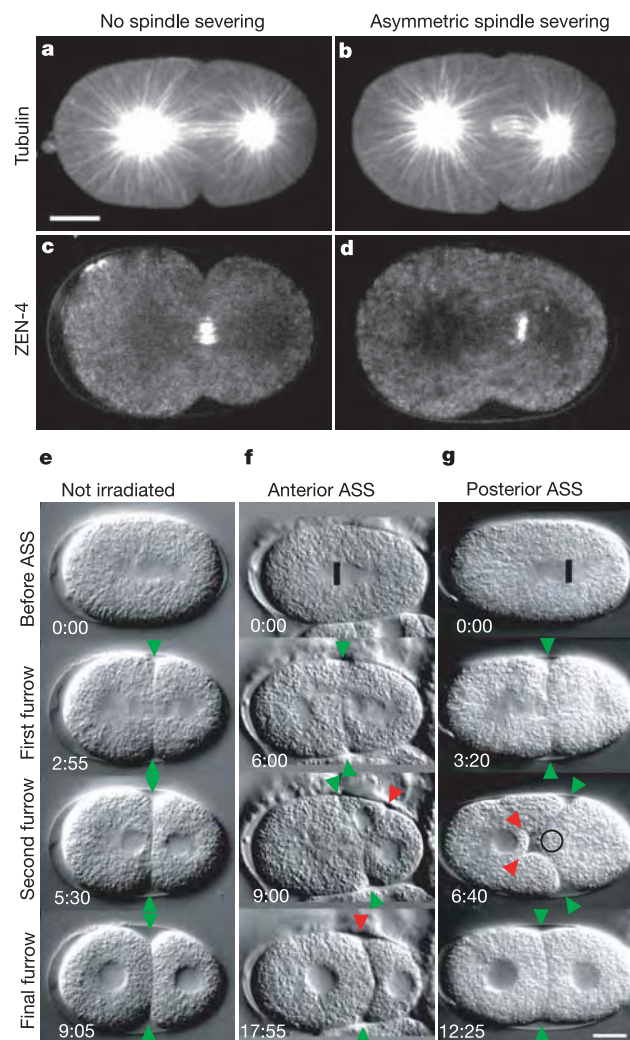


Figure 1 | Asymmetric spindle severing spatially separates the spindle midzone and the region midway between the asters, leading to the generation of two furrows. **a–d**, Alteration of spindle geometry. Spinning-disk microscopy images after ASS performed in a yellow fluorescent protein (YFP)- α -tubulin background (**a**, **b**) or a zen-4-GFP background (**c**, **d**) to reveal the microtubule cytoskeleton and the spindle midzone, respectively. Unsevered control cells are shown in **a** and **c**, severed spindles in **b** and **d**. **e–g**, Generation of two furrows after ASS shown in DIC microscopy image series (compare Supplementary Videos 1–5). Time shown in minutes following ASS procedure. **e**, Wild-type embryo. **f**, Separation of the anterior aster. **g**, Separation of the posterior aster. The irradiated region is indicated by a black bar. The posterior nucleus lies outside the focal plane in one image (position indicated by a black circle). Furrows often were unilateral after ASS. The asters seem to be attached to the ingressing furrow, which may cause the midzone to move off axis (see **f**). First furrows are indicated in green, second furrows in red. The difference in furrowing after posterior and anterior ASS is probably due to the geometry of the cell. After anterior ASS, the midzone moves closer to the cortex compared with posterior ASS. In general, cytokinesis takes longer after ASS (roughly double the time compared to control). Posterior is to the right in **a–g**. Scale bars, 10 μ m.

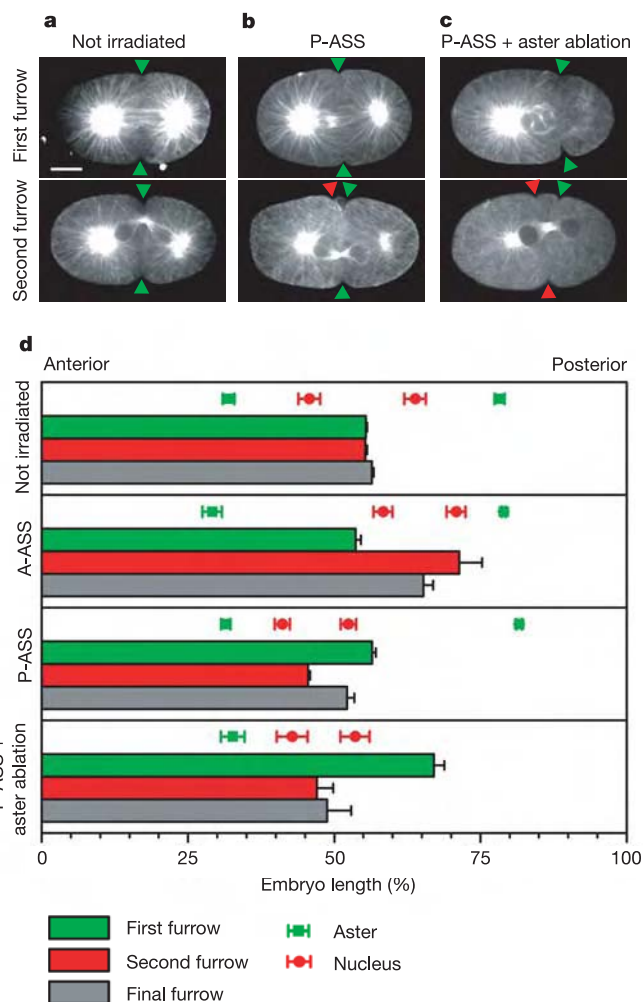


Figure 2 | Microtubule asters position the first cytokinesis furrow. The spindle midzone positions the second cytokinesis furrow. **a–c**, Spinning-disk microscopy images after posterior ASS (P-ASS) with (**c**) and without (**b**) subsequent aster ablation, observed by YFP- α -tubulin fluorescence. **a**, Unsevered control. Embryos in the top panels are shown at the point at which the first furrow (green arrowheads) ingresses. Compared with **a** and **b**, the furrow in **c** is set up further away from the remaining aster. Embryos in the bottom panels are shown at the point at which the second furrow ingresses. The second furrow (red arrowheads in **b** and **c**) always aims at the spindle midzone. Posterior is on the right. Scale bar, 10 μ m. Compare Supplementary Videos 6 and 7 showing aster ablation in DIC microscopy. **d**, Quantification of spindle dimensions and furrow positions in unsevered embryos and embryos after anterior (A) and posterior (P) ASS, and P-ASS plus aster ablation. The position of the aster centres (green squares) and the nuclei (red dots) at the time the first furrow ingresses are indicated. The positions of the first (green), second (red) and final (grey) furrow are shown as histograms. Error bars are s.e.m.

were capable of division but failed to segregate the DNA to the two daughter cells, yielding one cell with an aster and one cell with an aster and two nuclei. Progression of aster-positioned cleavage was fast compared with wild type ($6 \pm 1 \mu\text{m s}^{-1}$ in wild-type embryos and $11 \pm 1 \mu\text{m s}^{-1}$ in *spd-1(oj5)* embryos), and the furrow did not

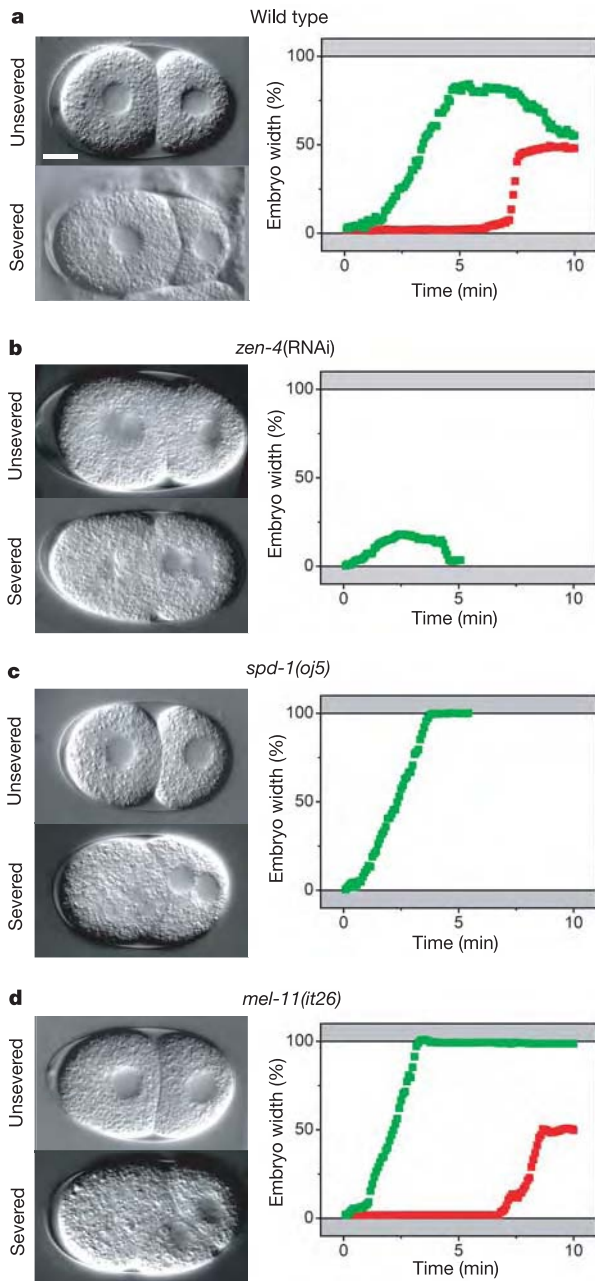


Figure 3 | Genetic control of redundant cleavage plane positioning. **a–d**, Phenotypic defects after ASS. Unsevered control and severed embryos are shown at the point at which the cytokinesis furrows are maximally ingressed. Right panels show cleavage progression over time, with the aster-positioned furrow in green and the midzone-positioned furrow in red. **a**, In wild-type embryos, the aster-positioned furrow does not complete, but instead pauses. The cytokinesis furrow position is then corrected by the midzone, and part of the aster-positioned furrow regresses. **b**, Both furrowing activities are affected, leading to a failure in cytokinesis. **c**, No correction mechanism is present; the embryo divides solely using its aster-based mechanism, resulting in one cell that contains no nucleus, and one cell that contains two nuclei. **d**, Failure to correct the aster-positioned furrow means that the aster-positioned furrow completes before the midzone-positioned furrow starts, leading to the generation of three cells.

pause, suggesting that the spindle midzone inhibits the aster-positioned furrow. The absence of a midzone-positioned furrow indicates that the defective spindle midzone in *spd-1* and *kfp-7* mutants is not able to provide a positional cue. Additionally, the completion of cytokinesis in *spd-1* and *kfp-7* mutants solely on the basis of an aster-positioned furrow demonstrates that cytokinesis can complete without a functional midzone, confirming previous studies^{12,20}. Cytokinesis in *spd-1* and *kfp-7* mutants highlights the redundancy of the cleavage furrow specification mechanism. This experiment further shows that the two furrow-positioning mechanisms are genetically separable.

The fourth class consisted of one gene, *mel-11*, which is a myosin light chain phosphatase that inhibits non-muscle myosin (NMY-2) contractility (Fig. 3d). In *mel-11* embryos, cytokinesis occurs twice as fast as in wild-type embryos²¹. In this class, the aster-positioned furrow ingressed and completed. However, a midzone-dependent furrow was also specified and completed, resulting three cells: one containing an aster, one containing a nucleus, and one containing an aster and a nucleus. We measured the progression of aster-dependent cleavage as $6 \pm 1 \mu\text{m s}^{-1}$ in wild-type embryos and $11 \pm 2 \mu\text{m s}^{-1}$ in *mel-11(it26)* embryos. Thus, reduced ingression of the aster-positioned furrow in the presence of a spindle midzone is probably regulated through myosin activity, and facilitates a positional correction mechanism by the spindle midzone.

By physically changing the geometry of the mitotic apparatus in a *C. elegans* embryo at the one-cell stage, we have shown that the position of the cytokinesis furrow is controlled by the consecutive and concerted action of two signals. The first positional signal is provided by microtubule asters. This is followed by a second signal specified by the spindle midzone. The first, aster-derived signal alone seems to be sufficient for cleavage. The second signal can probably also cleave the cell alone, as the midzone can position contractile material and induce a furrow after ASS. The second signal provides a robust correction mechanism to ensure that the genetic material is segregated equally to the two daughter cells. The spindle midzone is not required for completion and abscission in our experiments. However, the spindle midzone negatively influences the ingression of the aster-positioned furrow, probably by regulating myosin activity, perhaps through MEL-11. It facilitates the correction mechanism by the spindle midzone. The dominance of the midzone-positioned furrow is further highlighted by a persistent flow of green fluorescent protein (GFP)-tagged myosin (NMY-2-GFP) patches into the midzone-positioned furrow, while flow into the aster-positioned furrow ceases (Supplementary Videos 14 and 15). Redundant cytokinesis signals could be a universal mechanism to increase the precision with which the cleavage plane is positioned. Different cell types, however, might use the two mechanisms to different extents.

METHODS

Laser setup, microscopy and image analysis. The laser ablation experiments were performed using a highly focused ultraviolet laser beam and observed using differential interference contrast (DIC) and spinning-disk microscopy as described²², except that a Melles Griot Arlon Laser (488 nm, 100 mW) laser was used for fluorescence imaging. All ablations were performed in DIC optics. Quantification of spindle dimensions and furrow positions were made using Scion Image. Cytokinesis furrows were tracked in DIC movies using Metamorph software. All errors are reported as s.e.m. Figs were prepared using Photoshop and Illustrator (Adobe).

Asymmetric spindle severing (ASS). Asymmetric spindle severing was performed just after anaphase onset (identified by the disappearance of the metaphase plate) and observed by DIC microscopy. At the time of the spindle severing, spindle midzone components like ZEN-4 are already localized to the spindle midzone (unpublished observations using ZEN-4-GFP). Five to ten shots were fired at the region midway between one aster centre and the separated chromatin (see Supplementary Videos 1 and 2). One shot consisted of 25 optical pulses at a repetition rate of 250 Hz. If the same number of shots was fired in the region between spindle and cortex, no severing of the spindle is observed (data

not shown). The position of the first furrow was measured at the point at which it had ingressed roughly two-thirds of the egg width. The position of the second furrow was measured at the first point at which the furrow was visible. If the furrow showed bent growth, we measured the final position to which the furrow was aiming. The position of the final furrow was measured ($n = 4$) at the point at which the two blastomeres rounded up and the furrow was perpendicular to the anterior–posterior axis of the embryo. For furrow tracking ($n = 3$), we focused such that the largest opening of the contractile ring was visible. We compared the tracking data for an NMY-2–GFP embryo when collected by DIC or fluorescence microscopy and obtained comparable results. For simplicity, DIC videos were used for all quantifications. After ASS, the aster-positioned furrow is normally unilateral in wild-type embryos but bilateral in MEL-11 and SPD-1 embryos, which accounts for the different cytokinesis progression speeds.

Aster ablation. In order to ablate the asters, 15–20 additional shots were delivered in the centre of the aster (the region devoid of yolk granules) after it had been isolated from the mitotic spindle. The microscope stage was rapidly moved in circles around the centre of the aster during shooting to prevent spreading of the aster fragments (Supplementary Video 6). If the shots were distributed in the astral region containing yolk granules, the aster remained intact and the first furrow was not displaced (data not shown). Optical sections were taken with a spinning-disk microscope. We scanned through the embryo to verify the absence of the aster. Furrow positions were quantified (as for ASS) for five embryos. Using DIC and spinning-disk microscopy, we observed the cell containing the irradiated aster for a time equivalent to several additional cell division cycles, and did not see any sign of an aster reforming.

Worm culture, strains, alleles and RNAi. *C. elegans* were cultured as described⁹. All RNAi-mediated depletions were performed in the RNAi hypersensitive strains NL2099 (ref. 23) or GR1373 (ref. 24). Both strains behaved like wild-type N2 worms after ASS. RNAi was performed either by injection of double-stranded RNA or by feeding, as described^{16,25}. Worms were analysed 48 or 72 h after injection or feeding, respectively. For feeding experiments, worms were mated with males to maintain egg-laying. Mutants were shifted to a restrictive temperature 5–24 h before analysis. For a list of mutants and RNAi methods used, see Supplementary Table 1. Genes were classified as wild type if the embryos displayed the published phenotype and if, under those conditions, two furrows were observed. Typically, 5–10 experiments were performed for each gene. All DIC experiments were carried out at 23 °C, spinning-disk data were collected at 17 °C.

Received 7 February; accepted 16 May 2005.

- Rappaport, R. *Cytokinesis in Animal Cells* (Cambridge Univ. Press, Cambridge, 1996).
- Bonaccorsi, S., Giansanti, M. G. & Gatti, M. Spindle self-organization and cytokinesis during male meiosis in *asterless* mutants of *Drosophila melanogaster*. *J. Cell Biol.* **142**, 751–761 (1998).
- Cao, L. G. & Wang, Y. L. Signals from the spindle midzone are required for the stimulation of cytokinesis in cultured epithelial cells. *Mol. Biol. Cell* **7**, 225–232 (1996).
- Rappaport, R. & Rappaport, B. N. Surface contractile activity associated with isolated asters in cylindrical sand dollar eggs. *J. Exp. Zool.* **235**, 217–226 (1985).
- Rappaport, R. Experiments concerning the cleavage stimulus in sand dollar eggs. *J. Exp. Zool.* **148**, 81–89 (1961).
- Wang, Y. L. The mechanism of cytokinesis: reconsideration and reconciliation. *Cell Struct. Funct.* **26**, 633–638 (2001).
- Canman, J. C. & Wells, W. A. Rappaport furrows on our minds: the ASCB Cytokinesis Meeting Burlington, VT July 22–25, 2004. *J. Cell Biol.* **166**, 943–948 (2004).
- Dechant, R. & Glotzer, M. Centrosome separation and central spindle assembly act in redundant pathways that regulate microtubule density and trigger cleavage furrow formation. *Dev. Cell* **4**, 333–344 (2003).
- Grill, S. W., Gonczy, P., Stelzer, E. H. & Hyman, A. A. Polarity controls forces governing asymmetric spindle positioning in the *Caenorhabditis elegans* embryo. *Nature* **409**, 630–633 (2001).
- Mishima, M., Pavicic, V., Gruneberg, U., Nigg, E. A. & Glotzer, M. Cell cycle regulation of central spindle assembly. *Nature* **430**, 908–913 (2004).
- Rappaport, R. Repeated furrow formation from a single mitotic apparatus in cylindrical sand dollar eggs. *J. Exp. Zool.* **234**, 167–171 (1985).
- Hiramoto, Y. Cell division without mitotic apparatus in sea urchin eggs. *Exp. Cell Res.* **11**, 630–636 (1956).
- Canman, J. C. et al. Determining the position of the cell division plane. *Nature* **424**, 1074–1078 (2003).
- Shelton, C. A., Carter, J. C., Ellis, G. C. & Bowerman, B. The nonmuscle myosin regulatory light chain gene *mlc-4* is required for cytokinesis, anterior–posterior polarity, and body morphology during *Caenorhabditis elegans* embryogenesis. *J. Cell Biol.* **146**, 439–451 (1999).
- Severson, A. F., Baillie, D. L. & Bowerman, B. A Formin Homology protein and a profilin are required for cytokinesis and Arp2/3-independent assembly of cortical microfilaments in *C. elegans*. *Curr. Biol.* **12**, 2066–2075 (2002).
- Sonnichsen, B. et al. Full-genome RNAi profiling of early embryogenesis in *Caenorhabditis elegans*. *Nature* **434**, 462–469 (2005).
- Schumacher, J. M., Golden, A. & Donovan, P. J. AIR-2: An Aurora/Ipl1-related protein kinase associated with chromosomes and midbody microtubules is required for polar body extrusion and cytokinesis in *Caenorhabditis elegans* embryos. *J. Cell Biol.* **143**, 1635–1646 (1998).
- Raich, W. B., Moran, A. N., Rothman, J. H. & Hardin, J. Cytokinesis and midzone microtubule organization in *Caenorhabditis elegans* require the kinesin-like protein ZEN-4. *Mol. Biol. Cell* **9**, 2037–2049 (1998).
- Jantsch-Plunger, V. et al. CYK-4: A Rho family GTPase activating protein (GAP) required for central spindle formation and cytokinesis. *J. Cell Biol.* **149**, 1391–1404 (2000).
- Verbrugghe, K. J. & White, J. G. SPD-1 is required for the formation of the spindle midzone but is not essential for the completion of cytokinesis in *C. elegans* embryos. *Curr. Biol.* **14**, 1755–1760 (2004).
- Piekny, A. J. & Mains, P. E. Rho-binding kinase (LET-502) and myosin phosphatase (MEL-11) regulate cytokinesis in the early *Caenorhabditis elegans* embryo. *J. Cell Sci.* **115**, 2271–2282 (2002).
- Grill, S. W., Howard, J., Schaffer, E., Stelzer, E. H. & Hyman, A. A. The distribution of active force generators controls mitotic spindle position. *Science* **301**, 518–521 (2003).
- Simmer, F. et al. Loss of the putative RNA-directed RNA polymerase RRF-3 makes *C. elegans* hypersensitive to RNAi. *Curr. Biol.* **12**, 1317–1319 (2002).
- Kennedy, S., Wang, D. & Ruvkun, G. A conserved siRNA-degrading RNase negatively regulates RNA interference in *C. elegans*. *Nature* **427**, 645–649 (2004).
- Kamath, R. S. et al. Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* **421**, 231–237 (2003).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank the *Caenorhabditis* Genetics Center, the UK HGMP Resource Centre, Cenix Bioscience, T. Stiernagle, J. Ahringer, M. Glotzer, M. Srayko, C. Cowan, L. Pelletier, P. E. Mains, A. Schwager and A. Pozniakovskiy for reagents; C. Cowan, S. Grill, S. Schoenegg, J. Stamford and M. Srayko for advice and help on the experimental setup; and C. Cowan, J. Howard and K. Verbrugghe for comments on the manuscript. H.B. was supported by a predoctoral fellowship from the Boehringer Ingelheim Foundation.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to H.B. (bringman@mpi-cbg.de).

Transcription of mammalian messenger RNAs by a nuclear RNA polymerase of mitochondrial origin

Julia E. Kravchenko^{1,2}, Igor B. Rogozin³, Eugene V. Koonin³ & Peter M. Chumakov^{1,2}

Transcription of eukaryotic genes is performed by three nuclear RNA polymerases, of which RNA polymerase II is thought to be solely responsible for the synthesis of messenger RNAs¹. Here we show that transcription of some mRNAs in humans and rodents is mediated by a previously unknown single-polypeptide nuclear RNA polymerase (spRNAP-IV). spRNAP-IV is expressed from an alternative transcript of the mitochondrial RNA polymerase gene (*POLRMT*). The spRNAP-IV lacks 262 amino-terminal amino acids of mitochondrial RNA polymerase, including the mitochondrial-targeting signal, and localizes to the nucleus. Transcription by spRNAP-IV is resistant to the RNA polymerase II inhibitor α -amanitin but is sensitive to short interfering RNA specific for the *POLRMT* gene. The promoters for spRNAP-IV differ substantially from those used by RNA polymerase II, do not respond to transcriptional enhancers and contain a common functional sequence motif.

Replication and expression of the eukaryotic mitochondrial genome fully depends on the nuclear genome: during evolution, most of the genes derived from the DNA of the proto-mitochondrial endosymbiont and required for mitochondrial function were translocated to the nucleus, such that most components of the mitochondrial proteome are imported into the mitochondria^{2,3}. The mitochondrial RNA polymerase (mtRNAP) is a homologue of the bacteriophage single-polypeptide RNA polymerases and displaced the typical, multisubunit bacterial RNA polymerase during the evolution of the endosymbiont^{4,5}. The mtRNAP is encoded in the spliced 3.8-kilo-base transcript⁶ from the nuclear gene *POLRMT*. The human mtRNAP precursor has an N-terminal mitochondria-targeting peptide⁶, which is cleaved from the protein during mitochondrial import⁷. The N-terminal structure of mtRNAP is supported by 19 available expressed sequence tags (ESTs); however, three ESTs correspond to an alternative structure, with a longer exon 1 including the proximal 224 base pairs (bp) of intron 1. The open reading frame (ORF) encoded by the alternative transcript is truncated at the N terminus and would specify a protein lacking the mitochondria-targeting peptide.

To verify the existence and to explore potential functions of an alternative, possibly non-mitochondrial product of the *POLRMT* gene, we designed primers to show variations within the 5' region of the *POLRMT* transcripts. By sequencing polymerase chain reaction with reverse transcription (RT-PCR) products produced with these primers, we found that a fraction of the transcripts in human, mouse and rat cell lines correspond to the splice variant retaining intron 1 sequences (see Supplementary Fig. 1a, b), confirming that the transcripts specifying an N-terminally truncated mtRNAP are produced in both primates and rodents.

To determine whether the shorter polypeptide lacking the mitochondrial targeting peptide is indeed synthesized, we analysed proteins that react with antibodies to a carboxy-terminal peptide

of mtRNAP (ref. 8). Confocal microscopy revealed both mitochondrial and nuclear localization of immunoreactive proteins in HeLa cells (Fig. 1a). On western blots, total cell extracts showed two protein bands of ~135 kDa and ~110 kDa. In isolated mitochondria there was a single band of ~135 kDa, whereas nuclear extracts contained only the ~110 kDa band (Fig. 1b). Total extracts from the mitochondrial DNA-depleted ρ 0 HeLa cells⁹ (see Supplementary Fig. 2a) contained reduced amounts of the mitochondrial 135 kDa form, whereas the nuclear 110 kDa protein was unaffected (Fig. 1b). Inhibition of the *POLRMT* gene by RNA interference (RNAi) resulted in abrogation of both nuclear and mitochondrial immunostaining (see Supplementary Fig. 2b) and almost complete disappearance of the two protein forms (Fig. 1b) and both transcripts from the *POLRMT* gene (see Supplementary Fig. 1a). By contrast, RNAi against the proximal portion of intron 1 caused abrogation of nuclear immunostaining concomitant with the disappearance of the nuclear 110 kDa band, with no changes in mitochondria (Fig. 1d; see also Supplementary Fig. 2b).

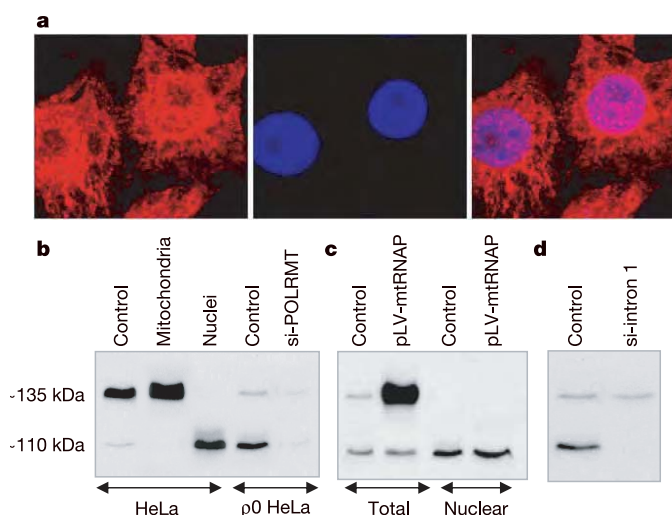


Figure 1 | The *POLRMT* gene is a source of mitochondrial and nuclear protein products. **a**, Confocal microscopy of HeLa cells stained with antibodies to *POLRMT* peptide and TRITC anti-rabbit IgG counterstained for nuclei with Hoechst 33258. **b**, Two *POLRMT* gene products detected by western blotting in HeLa (three left lanes) and ρ 0 HeLa cells (two right lanes) with antibodies to *POLRMT* peptide. Shown are total cell extracts (lanes 1, 4, 5), mitochondrial (lane 2) and nuclear (lane 3) fractions, and sample from cells expressing siRNA to exons 3 and 17 of human *POLRMT* gene (lane 5). **c**, Western blot showing the expression of the mtRNAP-encoding ORF in ρ 0 HeLa cells. **d**, Western blot with antibodies to the *POLRMT* peptide showing the expression of siRNA targeting *POLRMT* intron 1 in ρ 0 HeLa cells.

¹Lerner Research Institute, The Cleveland Clinic Foundation, Cleveland, Ohio 44195, USA. ²Engelhardt Institute of Molecular Biology, 119991 Moscow, Russia. ³National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA.

Overexpression of the complementary DNA encoding the full-length mtRNAP (ref. 6) caused a substantial increase in the mitochondrial 135-kDa band but no increase in the 110-kDa nuclear form (Fig. 1c; see also Supplementary Fig. 2b), suggesting that the latter protein does not originate from the mtRNAP mRNA neither through internal initiation of translation nor processing of the 135 kDa precursor. We then overexpressed the predicted alternative ORF and, separately, the mtRNAP ORF in $\rho 0$ HeLa cells, both tagged with a C-terminal Flag epitope. Immunostaining showed exclusively mitochondrial localization of the overexpressed mtRNAP and exclusively nuclear localization of the protein encoded by the alternative ORF (Fig. 2a). The electrophoretic mobility of the overexpressed Flag-tagged proteins corresponded precisely to the two endogenous bands as shown with POLRMT-specific antibodies (Fig. 2b), although the size of the nuclear protein determined by electrophoresis (~ 110 kDa) was much smaller than the size of the alternative ORF product predicted from the mRNA sequence (~ 130 kDa). Attaching an N-terminal Flag to the alternative ORF resulted in a protein that did not react with anti-Flag antibodies but reacted with antibodies to mtRNAP (Fig. 2c), suggesting that the first ATG was not used for the translation initiation. Indeed, the product of an N-terminally truncated alternative ORF fragment starting at the first initiation codon conserved between mouse and human (ATG-6) was identical in size to the endogenous nuclear protein, whereas the fragment starting at ATG-8 produced a smaller protein (see Supplementary Figs 3 and 4). Thus, the nuclear form of the protein seems to start at ATG-6 of the ORF and correspond to the 968 C-terminal amino acids of mtRNAP, which include the entire catalytic domain (see Supplementary Fig. 3).

Our findings suggest that the *POLRMT* gene specifies two single-polypeptide polymerases, one (mtRNAP) targeted to the mitochondria, and the other (spRNAP-IV) having a putative function in the nucleus. Conceivably, spRNAP-IV could be the fourth nuclear RNA polymerase of animal cells, although alternative functions of this

protein, other than direct involvement in transcription, can not be ruled out. To determine whether or not spRNAP-IV is important for cell viability, we inhibited the expression of *POLRMT* gene by siRNA expression. This resulted in a substantial growth retardation and cell death, which was reversed by co-expression of recombinant spRNAP-IV (Fig. 2d). Therefore, spRNAP-IV seems to have a distinct role in cell physiology.

In an attempt to identify genes that could be transcribed by the putative spRNAP-IV, we hypothesized that transcription of these genes would be resistant to α -amanitin, a specific inhibitor of RNA polymerase II (RNAP-II). Hybridization of a total RNA probe from HeLa cells treated with α -amanitin with an Affymetrix microarray U133A showed ~ 70 upregulated (twofold or greater) transcripts. Of the 15 transcripts tested by RT-PCR, four showed lack of sensitivity to or even stimulation by α -amanitin (see Supplementary Fig. 5a, b). A similar pattern was observed for the orthologous transcripts in mouse 3T3 cells (Supplementary Fig. 5c). In northern hybridization, the α -amanitin-resistant transcripts encoding the zinc-finger BTB domain-containing protein 1 (*ZBTB1*), prenylcysteine oxidase (*MGC3265*) and aldehyde dehydrogenase 8A1 (*ALDH12*) genes showed dose-dependent stimulation by α -amanitin, whereas control transcripts from the *p21^{WAF1/CIP1}* (*CDKN1A*) and *GAPDH* genes were sensitive to α -amanitin (Fig. 3a; see also Supplementary Fig. 6). When RNAP-I or RNAP-II was inhibited by RNAi, gradual decrease and disappearance of transcripts encoding each of these polymerases was observed by day 4, which correlated with substantial stimulation of *ZBTB1*, *MGC3265* and *ALDH12* transcripts (Fig. 3b). Inhibition of the largest subunit of RNAP-III with siRNA oligonucleotides also resulted in slightly increased levels of all three transcripts (see Supplementary Fig. 6b). Thus, these genes apparently are not transcribed by known nuclear RNA polymerases. By contrast, inhibition of *POLRMT* transcripts in $\rho 0$ HeLa cells by RNAi was accompanied by a dramatic decrease in transcripts from the α -amanitin-resistant genes (Fig. 3c), which was reversed by overexpression of the spRNAP-IV

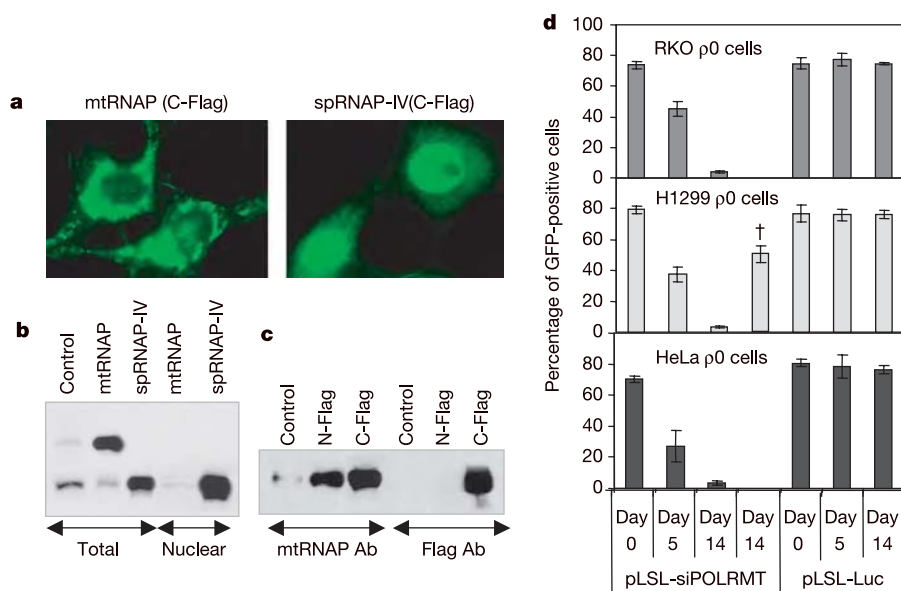


Figure 2 | Two distinct *POLRMT*-specific polypeptides are encoded by alternative transcripts. **a**, Immunostaining with anti-Flag antibodies of HeLa cells transduced with Flag-tagged ORF for mtRNAP (left panel) or Flag-tagged alternative ORF encoding spRNAP-IV (right panel). **b**, Western blot analysis of proteins reactive with antibodies to a *POLRMT* peptide after overexpression of mtRNAP ORF (lanes 2 and 4) or spRNAP-IV ORF (lanes 3 and 5). Total cell extracts (lanes 2 and 3) or nuclei (lanes 4 and 5) are shown. **c**, Western blot analysis of HeLa cells transduced with spRNAP-IV Flag-tagged at the N- or C terminus and stained with antibodies to *POLRMT*

(left) or to Flag (M2). **d**, Expression of *POLRMT* siRNA affects cell growth. Specified $\rho 0$ cell lines were infected with lentivirus vector LSLG co-expressing GFP and siRNA targeting *POLRMT* exon 3 (or exon 17, which produced a similar result), or GFP and siRNA targeting luciferase. The infected cells were mixed with uninfected cells, and the proportion of fluorescent cells was monitored by FACS. A control sample of H1299 $\rho 0$ HeLa cells was co-infected with lentivirus vector expressing spRNAP-IV ORF (indicated by a dagger). Data represent the means of two experiments with standard deviation (s.d.).

ORF (Fig. 3d). These results indicate that the truncated nuclear protein product of the *POLRMT* gene participates in the production of certain mRNAs.

To examine the repertoire of transcripts affected by spRNAP-IV, we monitored changes in the expression pattern in $\rho 0$ HeLa cells after 48 h expression of siRNA targeting the *POLRMT* gene using hybridization to a cDNA microarray. Of $\sim 20,000$ transcripts that gave a positive signal, $\sim 1,000$ transcripts were substantially inhibited

(twofold or greater), whereas ~ 300 transcripts showed comparable levels of upregulation (see Supplementary Table 1). Certainly, some of the changes might be attributed to secondary effects of inhibition of the true spRNAP-IV-regulated genes. Also, cooperation between spRNAP-IV and RNAP-II in the expression of certain genes cannot be ruled out. Of 17 randomly selected transcripts that showed down-regulation on the microarray, nine demonstrated an α -amanitin-resistant/inducible expression when assayed by RT-PCR (data not shown), suggesting that the number of mammalian spRNAP-IV-dependent genes might be in the hundreds. Upregulation of some of these transcripts by α -amanitin was confirmed by northern analysis in three different $\rho 0$ HeLa cell lines (see Supplementary Fig. 7).

To analyse promoters used for initiation of transcription by spRNAP-IV, we amplified and cloned the upstream regions of *ALDH12* and *ZBTB1* genes into a reporter plasmid. When they were transfected into HeLa cells, both putative spRNAP-IV promoters supported luciferase expression with activities comparable to that of the SV40 early promoter (Fig. 4a). The expression of luciferase driven by the *ALDH12* and *ZBTB1* promoters was resistant to α -amanitin, in contrast to the α -amanitin-sensitive expression from the SV40 promoter (Fig. 4a). Inhibition of *POLRMT* by RNAi 24 h before the transfection of reporter constructs led to a decreased activity of the *ALDH12* promoter but did not affect the SV40 promoter (Fig. 4b). To determine whether spRNAP-IV binds to the *ALDH12* promoter, formaldehyde-fixed extracts of HeLa cells transfected with an empty vector or with the constructs bearing *ALDH12* and SV40 promoters were precipitated with antibodies to RNAP-II or *POLRMT*, and the number of ampicillin-resistant colonies indicating the efficiency of immunoprecipitation was determined by transformation of *Escherichia coli*. For the *ALDH12*-promoter-containing plasmid, substantial enrichment was observed with the *POLRMT* antibodies; a tenfold lower enrichment was seen with antibodies to RNAP-II, suggesting that spRNAP-IV preferentially binds to this construct. The reciprocal enrichment was observed with the SV40 promoter construct (Fig. 4c).

A comparison of the upstream sequences of the *ALDH12*, *ZBTB1* and *MGC3265* genes revealed two shared motifs (see Supplementary Fig. 8); counterparts to these motifs were also identified in several additional genes putatively transcribed by spRNAP-IV, including *ULK1* (data not shown). Both motifs were conserved among the orthologous genes from other mammals (see Supplementary Fig. 8). The putative spRNAP-IV promoter motifs showed no detectable similarity to the sequences of RNAP I–III promoters^{10,11}, those of the known transcription-factor-binding sites¹², the sequences of mtDNA that are involved in mtRNAP-directed transcription, or the promoters of T-odd bacteriophage genes transcribed by single-polypeptide phage polymerases homologous to mtRNAP. Deletion of the sequences spanning these motifs from the *ALDH12*, *ZBTB1* and *MGC3265* promoters abrogated α -amanitin-resistant expression of luciferase, suggesting that the motifs are, indeed, functional elements of spRNAP-IV promoters (see Supplementary Fig. 9).

Transcription initiated by RNAP-II can be substantially activated by transcription factors that bind to transcriptional enhancers¹. We tested whether the activity of promoters of spRNAP-IV-controlled genes could be stimulated by placing a strong cytomegalovirus (CMV) enhancer close to *ALDH12*, SV40 and mitochondrial DNA H-strand (mtHs) promoters in a luciferase reporter plasmid. On transfection into HeLa cells, the CMV enhancer stimulated α -amanitin-sensitive transcription from the SV40 promoter by more than tenfold. By contrast, the *ALDH12* promoter fragment did not respond to the enhancer at all, and its basal activity remained responsive to stimulation by α -amanitin. Similarly, in HeLa cells producing the artificial transcription factor tTA¹³, the activity of the *ALDH12* promoter did not change when coupled with the tetracycline-responsive element (see Supplementary Fig. 10), confirming that RNAP-II enhancers do not affect transcription by spRNAP-IV. The mtHs promoter was inactive in this system, suggesting that the

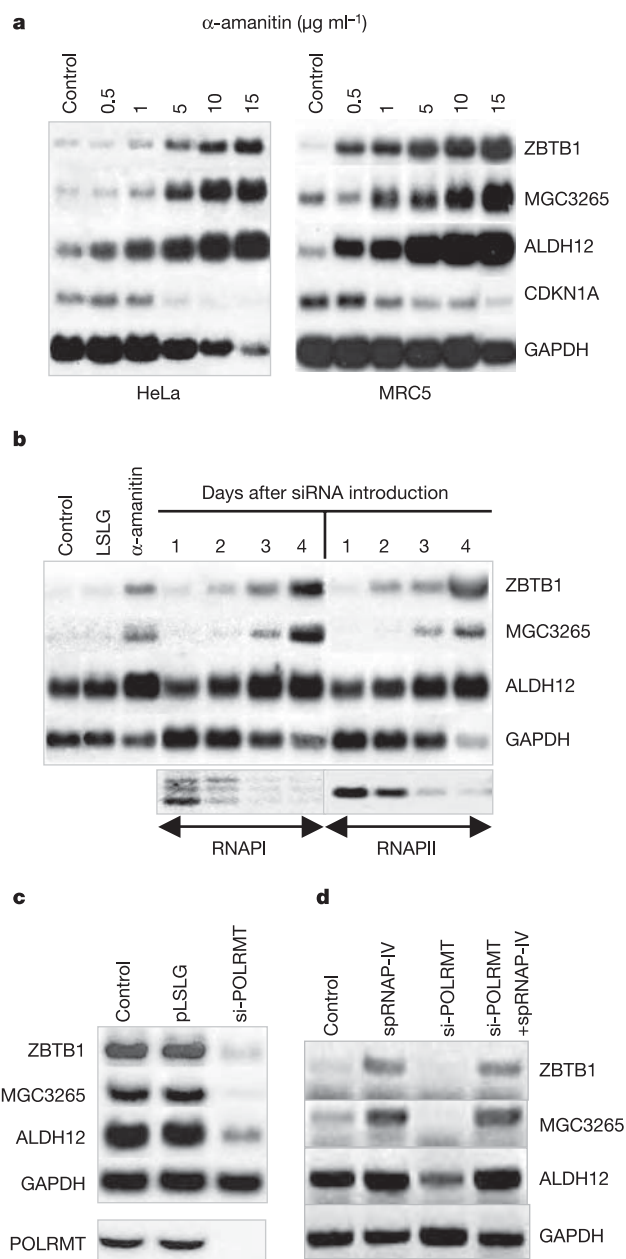


Figure 3 | A nuclear protein product of the *POLRMT* gene participates in the expression of the *ZBTB1*, *MGC3265* and *ALDH12* genes. a, Northern blot showing the dose-dependent stimulation of *ZBTB1*, *MGC3265* and *ALDH12* transcripts in HeLa and MRC5 cells treated for 14 h with α -amanitin. **b**, Northern blot analysis showing time-dependent stimulation of the transcripts following infection of HeLa cells with lentiviral vector LSLG expressing siRNAs targeting RNAP I or RNAP II. Inhibition of RNAP I and II mRNAs is shown on the lower panels. **c**, Northern blot showing the inhibition of the transcripts 4 days after infection of $\rho 0$ HeLa cells with LSLG lentivirus expressing siRNA targeting *POLRMT*. **d**, Northern blot showing the stimulation of the same transcripts after infection with a lentiviral vector expressing spRNAP-IV ORF (RT-PCR).

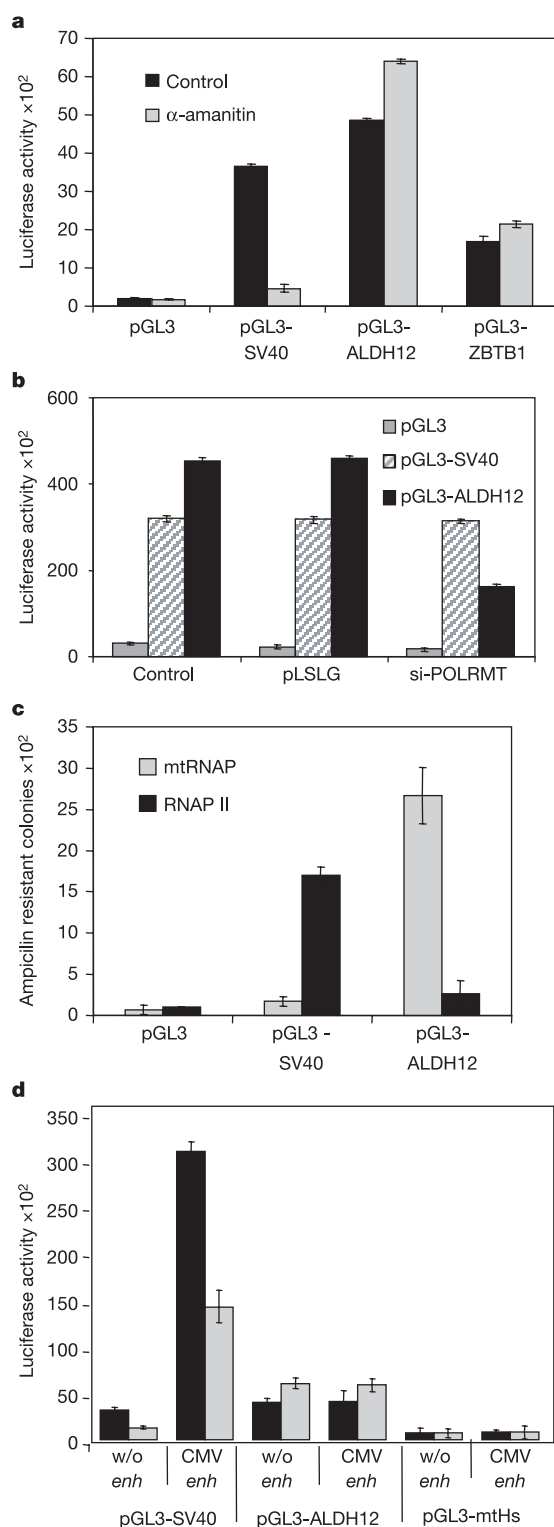


Figure 4 | Transcription from the *ALDH12* and *ZBTB1* promoters depends on spRNAP-IV. **a**, Luciferase assay measuring pGL3-luc constructs containing *ALDH12* (−482 to −12) and *ZBTB1* (−330 to +12) promoters that were transfected into HeLa cells (48 h). The cultures were treated with α -amanitin for 12 h. Data represent the means of two experiments with s.d. **b**, Suppression of *ALDH12* promoter by siRNA to *POLRMT*. Twenty-four hours after infection of HeLa cells with siRNA-expressing LSLG constructs the cells were transfected with pGL3-*ALDH12* (−482 to −12 upstream fragment) or control plasmids and assayed for luciferase activity over 48 h. Data represent the means of two experiments with s.d. **c**, Chromatin immunoprecipitation of promoter-containing plasmids transfected into HeLa cells with antibodies to RNAP II or *POLRMT*. The plasmids contained either *ALDH12* (−290 to −12) or SV40 promoters. Four experiments were averaged. Error bars indicate s.d. **d**, CMV-enhancer (−600 to −112, CMV enh) does not cooperate with *ALDH12* promoter (−290 to −12). Forty-eight hours after transfection of the promoter-containing pGL3-luc constructs (*ALDH12*, mtH-strand promoter and SV40 early promoter) the cultures were treated with α -amanitin for 12 h and assayed for luciferase. Data represent the means of two experiments with s.d.

participates in gene silencing, similarly to the recently identified structurally unrelated RNAP IV of *Arabidopsis*^{14,15}. A further important question is what are the transcription factors that affect spRNAP-IV function? The fact that spRNAP-IV promoters share no sequence similarity with mtDNA and that three known transcription factors (mtTFA, mtTFB1 and mtTFB2) that cooperate with mtRNAP seem to localize exclusively to the mitochondria^{16–18} suggest that spRNAP-IV might use a distinct set of factors for transcription of nuclear genes.

METHODS

Cell culture. Human carcinoma cell lines HeLa, A431, H1299 and RKO, normal human lung fibroblasts MRC5, mouse fibroblasts 3T3 and rat fibroblasts Rat1 were cultured in Dulbecco's modified Eagle medium (DMEM), supplemented with 10% fetal bovine serum (HyClone) and penicillin/streptomycin. Mitochondrial DNA-deficient ρ 0 HeLa were obtained, subcloned and cultured as described¹⁹. The absence of mtDNA was confirmed by PCR. The mitochondria in control and ρ 0 cells were stained with MitoTracker Red 589 (Invitrogen).

Northern analysis. Radioactive probes were prepared from DNA inserts of appropriate I.M.A.G.E. Consortium clones: *RNAP I* large subunit, I.M.A.G.E.: 4413694; *RNAP II* large subunit, I.M.A.G.E.: 4359716; *POLRMT*, I.M.A.G.E.: 6572256; *GAPDH*, MGC:2031, I.M.A.G.E.: 3543589; *p21* (*CDKN1A*), MGC:97077, I.M.A.G.E.:7262289; *ALDH12*, I.M.A.G.E.: 4998434; MGC3265, I.M.A.G.E.: 3506623; *ZBTB1*, MGC:60335, I.M.A.G.E.: 6141266.

RT-PCR amplification. Total RNA that was isolated by Trizol reagent was converted to cDNA using SuperScript II reverse transcriptase (Invitrogen). PCR amplifications (typically 20–25 cycles) were performed with primers specified in the figure legends.

Recombinant constructs. Inhibition of expression of different genes was achieved by infection with recombinant lentiviral vector constructs (pLSLG) expressing hairpin siRNAs controlled by H1 RNA gene promoter as previously described²⁰. The following 19–21-bp regions corresponding to appropriate mRNAs were present in the hairpin transcripts: *RNAP I* large subunit, 5'-GTTTACAAGCCACTGAATCGC-3' and 5'-GAAACCAGCTTCCAGTTTCTG-3'; *RNAP II* large subunit, 5'-GAACTATTCTCCAACAGTCC-3' and 5'-GATACACACCACAGTCTCCAA-3'; exon 3 of *POLRMT* gene, 5'-CAAAGATACCTGAGAAGGATA-3'; exon 17 of *POLRMT* gene, 5'-CAACACACGTAAAGCAGAAGAA-3'; intron 1 of *POLRMT* gene, 5'-GGCAAAGAAGGTAACACAA-3'. For the expression of recombinant mtRNAP cDNA⁶ (a gift from V. Tiranti) or truncated forms of mtRNAP we used modified lentiviral expression vector pLV (ref. 21) (from I. Verma). Preparation of high-titre lentiviral stocks and infection of target cell cultures were performed as described²¹.

Chromatin immunoprecipitation. pGL3-luc (Promega) constructs containing the 278-bp upstream region of human *ALDH12* gene (−290 to −12) or the 202-bp SV40 early promoter were lipofected into HeLa cells. The cells were fixed with 1% formaldehyde 24 h thereafter and the prepared cell extracts²² were incubated overnight with antibodies to either mtRNAP (1:200) or RNAP II (N-2, sc-899, Santa Cruz Biotech) and Protein G Agarose beads. The beads were washed six times with 20 mM Tris-HCl, 200 mM NaCl, 0.5% Triton-X100, 0.05% deoxycholate, 0.1% NP-40, 1 mM phenylmethylsulphonyl fluoride (PMSF)

regulation of spRNAP-IV transcription in the nucleus is distinct from that of the mtRNAP in the mitochondria (Fig. 4c).

Identification of spRNAP-IV as a second polymerase transcribing mRNAs in mammals raises a wealth of questions for future investigation. In particular, it remains to be determined at what stage of eukaryotic evolution was a single-polypeptide RNA polymerase recruited for a nuclear role, what is the repertoire of genes transcribed by spRNAP-IV, how is the spRNAP-IV-dependent transcription regulated, how RNAP-II and spRNAP-IV divide and/or share their tasks in the transcription of mRNAs, and whether spRNAP-IV

containing 250 mM LiCl, and precipitates were eluted with 1% SDS and 200 mM NaCl. The samples were heated at 65 °C for 6 h, extracted with phenol/chloroform, ethanol-precipitated and transfected as five independent parallel samples into competent *DH5 α* *E. coli*. The number of ampicillin-resistant colonies representing immunoprecipitated plasmids was counted.

Microarray hybridizations. The microarray hybridizations were performed using total RNA from HeLa or $\rho 0$ HeLa cells treated with 10 $\mu\text{g ml}^{-1}$ α -amanitin for either 12 h or 48 h after introduction of siRNA-expressing lentivirus vector. The U133A or U133 2.0 Plus human cDNA microarrays were used at the Gene Expression Array Core Facility (Case Western Reserve University) according to Affymetrix protocols.

Western analysis. Samples of total protein, mitochondrial and nuclear fractions were isolated as described²³, heated at 95 °C, separated in 4–12% SDS polyacrylamide electrophoresis (SDS–PAGE), transferred to nitrocellulose membranes and probed with antibodies raised to a peptide (vnlepsdvpqdv) of human mtRNAP (gift from G. Shadel) or anti-Flag antibodies M2 (Sigma–Aldrich). The filters were developed with peroxidase-labelled anti-rabbit polyclonal antibodies and ECL-Plus Western detection reagents (Amersham Biosciences).

Immunostaining. Cells were fixed with 4% formaldehyde in phosphate-buffered saline (PBS) for 20 min, washed and incubated overnight at 4 °C with anti-mtRNAP or anti-Flag M2 antibodies (1:200) in blocking buffer²³. TRITC-labelled anti-rabbit or FITC-labelled anti-mouse (FITC) antibodies (Jackson Immune Research) were used for immunostaining.

Luciferase assays for promoter activity. Upstream regions of *ALDH12*, *ZBTB1* and *MGC3265* genes and mitochondrial DNA H-strand promoter (mtHs) were cloned into pGL3-luc reporter vector (Promega). The plasmids were transfected into HeLa or $\rho 0$ HeLa cells using Lipofectamine reagent (Invitrogen), along with a plasmid expressing β -galactosidase for normalization. The activity of luciferase was measured in cell extracts with the Luciferase assay system (Promega) using a fluorimetric plate reader.

Computational methods. Amino acid sequence alignments were constructed using the MACAW program²⁴. The search for nucleotide motifs in promoter regions and statistical assessment of the detected motifs was performed using the ParaMEME²⁵, Gibbs sampling²⁶, CONSENSUS²⁷ and MULTIPROFILER²⁸ methods. Search for known transcription-factor binding sites was performed using the MATCH program²⁹. The conservation of the promoter sequences in orthologous mammalian genes was determined using the BLAT server³⁰.

Received 13 February; accepted 19 May 2005.

- Lewin, B. *Genes VIII* 597–626 (Pearson Prentice Hall, New Jersey, 2004).
- Kurland, C. G. & Andersson, S. G. Origin and evolution of the mitochondrial proteome. *Microbiol. Mol. Biol. Rev.* **64**, 786–820 (2000).
- Gabalton, T. & Huynen, M. A. Reconstruction of the proto-mitochondrial metabolism. *Science* **301**, 609 (2003).
- Masters, B. S., Stohl, L. L. & Clayton, D. A. Yeast mitochondrial RNA polymerase is homologous to those encoded by bacteriophages T3 and T7. *Cell* **51**, 89–99 (1987).
- Gray, M. W. & Lang, B. F. Transcription in chloroplasts and mitochondria: a tale of two polymerases. *Trends Microbiol.* **6**, 1–3 (1998).
- Tiranti, V. *et al.* Identification of the gene encoding the human mitochondrial RNA polymerase (h-mtRPOL) by cyberscreening of the Expressed Sequence Tags database. *Hum. Mol. Genet.* **6**, 615–625 (1997).
- Taroni, F. & Rosenberg, L. E. The precursor of the biotin-binding subunit of mammalian propionyl-CoA carboxylase can be translocated into mitochondria as apo- or holoprotein. *J. Biol. Chem.* **266**, 13267–13271 (1991).
- Seidel-Rogol, B. L. & Shadel, G. S. Modulation of mitochondrial transcription in response to mtDNA depletion and repletion in HeLa cells. *Nucleic Acids Res.* **30**, 1929–1934 (2002).
- King, M. P. & Attardi, G. Injection of mitochondria into human cells leads to a rapid replacement of the endogenous mitochondrial DNA. *Cell* **52**, 811–819 (1988).
- Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annu. Rev. Biochem.* **72**, 449–479 (2003).
- Paule, M. R. & White, R. J. Survey and summary: transcription by RNA polymerases I and III. *Nucleic Acids Res.* **28**, 1283–1298 (2000).
- Matys, V. *et al.* TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**, 374–378 (2003).
- Gossen, M. & Bujard, H. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. *Proc. Natl Acad. Sci. USA* **89**, 5547–5551 (1992).
- Onodera, Y. *et al.* Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell* **120**, 613–622 (2005).
- Herr, A. J., Jensen, M. B., Dalmay, T. & Baulcombe, D. C. RNA polymerase IV directs silencing of endogenous DNA. *Science* **308**, 118–120 (2005).
- Rantanen, A., Jansson, M., Oldfors, A. & Larsson, N. G. Downregulation of Tfam and mtDNA copy number during mammalian spermatogenesis. *Mamm. Genome* **12**, 787–792 (2001).
- Falkenberg, M. *et al.* Mitochondrial transcription factors B1 and B2 activate transcription of human mtDNA. *Nature Genet.* **31**, 289–294 (2002).
- McCulloch, V. & Shadel, G. S. Human mitochondrial transcription factor B1 interacts with the C-terminal activation region of h-mtTFA and stimulates transcription independently of its RNA methyltransferase activity. *Mol. Cell. Biol.* **23**, 5816–5824 (2003).
- King, M. P. & Attardi, G. Isolation of human cell lines lacking mitochondrial DNA. *Methods Enzymol.* **264**, 304–313 (1996).
- Budanov, A. V., Sablina, A. A., Feinstein, E., Koonin, E. V. & Chumakov, P. M. Regeneration of peroxiredoxins by p53-regulated sestrins, homologs of bacterial AhpD. *Science* **304**, 596–600 (2004).
- Pfeifer, A., Ikawa, M., Dayn, Y. & Verma, I. M. Transgenesis by lentiviral vectors: lack of gene silencing in mammalian embryonic stem cells and preimplantation embryos. *Proc. Natl Acad. Sci. USA* **99**, 2140–2145 (2002).
- Kaesser, M. D. & Iggo, R. D. Chromatin immunoprecipitation analysis fails to support the latency model for regulation of p53 DNA binding activity *in vivo*. *Proc. Natl Acad. Sci. USA* **99**, 95–100 (2002).
- Harlow, E. & Lane, D. *Using Antibodies: A Laboratory Manual* 276–301; 103–148 (Cold Spring Harbor Laboratory Press, New York, 1999).
- Schuler, G. D., Altschul, S. F. & Lipman, D. J. A workbench for multiple alignment construction and analysis. *Proteins* **9**, 180–190 (1991).
- Grundy, W. N., Bailey, T. L. & Elkan, C. P. ParaMEME: a parallel implementation and a web interface for a DNA and protein motif discovery tool. *Comput. Appl. Biosci.* **12**, 303–310 (1996).
- Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
- Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563–577 (1999).
- Keich, U. & Pevzner, P. A. Subtle motifs: defining the limits of motif finding algorithms. *Bioinformatics* **18**, 1382–1390 (2002).
- Kel, A. E. *et al.* MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **31**, 3576–3579 (2003).
- Kent, W. J. BLAT—the BLAST-like alignment tool. *Genome Res.* **12**, 656–664 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The work was supported by funds provided by The Lerner Research Institute to P.M.C. and by NIH and RFBR to P.M.C.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to P.M.C. (chumakp@ccf.org).

LETTERS

TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing

Thimmaiah P. Chendrimada^{1*}, Richard I. Gregory^{1*}, Easwari Kumaraswamy^{1*†}, Jessica Norman¹, Neil Cooch¹, Kazuko Nishikura¹ & Ramin Shiekhattar¹

MicroRNAs (miRNAs) are generated by a two-step processing pathway to yield RNA molecules of approximately 22 nucleotides that negatively regulate target gene expression at the post-transcriptional level¹. Primary miRNAs are processed to precursor miRNAs (pre-miRNAs) by the Microprocessor complex^{2–4}. These pre-miRNAs are cleaved by the RNase III Dicer^{5–8} to generate mature miRNAs that direct the RNA-induced silencing complex (RISC) to messenger RNAs with complementary sequence⁹. Here we show that TRBP (the human immunodeficiency virus transactivating response RNA-binding protein¹⁰), which contains three double-stranded, RNA-binding domains, is an integral component of a Dicer-containing complex. Biochemical analysis of TRBP-containing complexes revealed the association of Dicer–TRBP with Argonaute 2 (Ago2)^{11,12}, the catalytic engine of RISC. The physical association of Dicer–TRBP and Ago2 was confirmed after the isolation of the ternary complex using Flag-tagged Ago2 cell lines. *In vitro* reconstitution assays demonstrated that TRBP is required for the recruitment of Ago2 to the small interfering RNA (siRNA) bound by Dicer. Knockdown of TRBP results in destabilization of Dicer and a consequent loss of miRNA biogenesis. Finally, depletion of the Dicer–TRBP complex via exogenously introduced siRNAs diminished RISC-mediated reporter gene silencing. These results support a role of the Dicer–TRBP complex not only in miRNA processing but also as a platform for RISC assembly.

To gain an insight into the components of the miRNA/siRNA processing machinery, we isolated a Dicer-containing complex from human cells. This was accomplished by developing HEK 293-derived stable cell lines expressing Dicer tagged with Flag (Flag-Dicer). Flag-Dicer was isolated using affinity chromatography, and the affinity eluate was subjected to SDS–polyacrylamide gel electrophoresis (PAGE) followed by silver staining and western blot analysis. Western blot and mass spectrometric analyses indicated that most polypeptides in the Dicer affinity eluate were products of the proteolytic break down of Dicer (Fig. 1a). However, mass spectroscopy identified a 50-kDa band (six peptide sequences that migrated slightly above the contaminating MEP50 band) corresponding to the human immunodeficiency virus (HIV)-1 transactivating response (TAR) RNA-binding protein (TRBP)¹⁰. The TRBP gene encodes a protein with three double-stranded RNA-binding domains (dsRBDs). Analysis of the non-redundant protein database by Blast identified proteins with close homology to TRBP in both vertebrates and *Drosophila* (CG6866) (Fig. 1b).

To confirm the association of TRBP and Dicer, we developed Flag-TRBP cell lines. Polypeptides associated with TRBP were isolated and subjected to chromatographic fractionation. Analysis of gel filtration fractions by silver staining and western blot analysis using antibodies

against Dicer and Flag epitope demonstrated the co-elution of a fraction of TRBP and Dicer as a component of an approximately 500-kDa complex (Fig. 1c). The presence of Dicer was also confirmed by mass spectrometric sequencing. Moreover, additional bands (indicated with an asterisk in Fig. 1c) correspond to SKB1 and MEP50, common contaminants of Flag purification. Although most of TRBP eluted in smaller fractions (32 and beyond; perhaps as a consequence of overexpression), a minor portion of TRBP eluted as a large complex (fractions 16 and 18) not easily visualized by silver staining. To demonstrate the direct association of Dicer and TRBP, we generated recombinant Dicer and TRBP proteins from insect cells and bacteria, respectively. Recombinant Dicer protein was further purified to near homogeneity by gel filtration (Fig. 1d, left panel). Next, we mixed stoichiometric amounts of recombinant Dicer and TRBP and analysed the elution profile of the complex by gel filtration. Consistent with the behaviour of the endogenous complex, recombinant Dicer–TRBP co-elute as a complex, peaking in fractions 28–30 (Fig. 1d, right panel). To confirm the direct physical association of TRBP with Dicer, and to rule out the possibility that the observed co-elution was due to overlapping elution profiles of the individual proteins, we examined the elution profile of TRBP alone by gel filtration. Analysis of Superose 6 column fractions by silver staining revealed that recombinant TRBP eluted in a distinct fraction (fraction 32) compared with that of the Dicer–TRBP complex (compare Fig. 1e and d).

We have shown previously that recombinant Droscha requires its binding partner, the dsRBD protein DGCR8, for processing primary miRNAs to pre-miRNAs². Unlike Droscha, recombinant Dicer was reported to process long double-stranded RNA to yield products of 21–23 nucleotides¹³. To assess the processing of pre-miRNA by the Dicer–TRBP complex, an approximately 60-nucleotide pre-miRNA corresponding to miR-(23a ~ 27a ~ 24-2) was prepared through the action of the Microprocessor complex. Addition of equal amounts of either Dicer alone or Dicer–TRBP to the reaction generated the final 22-nucleotide mature miRNA (Fig. 2a, left panel). Analysis of the processing of a number of miRNAs by different concentrations of Dicer or the Dicer–TRBP complex revealed a specific activity that was comparable for the two enzyme preparations (data not shown). We next monitored the processing of long double-stranded RNA by Dicer and Dicer–TRBP. Similar to results obtained with pre-miRNA, the addition of equal amounts of recombinant Dicer or recombinant Dicer–TRBP complex to the reaction yielded comparable amounts of cleavage products of approximately 22 nucleotides (Fig. 2a, right panel). These results are reminiscent of the association of the double-stranded RNA-binding protein R2D2 with *Drosophila* Dicer2. Indeed, in the absence of R2D2, Dicer2 can still process double-stranded RNA efficiently,

¹The Wistar Institute, 3601 Spruce Street, Philadelphia, Pennsylvania 19104, USA. [†]Present address: The Children's Hospital of Philadelphia, 34th Street and Civic Center Boulevard, Philadelphia, Pennsylvania 19014, USA.

*These authors contributed equally to this work.

but the resulting siRNAs are not effectively channelled into RISC¹⁴. Moreover, a recent report suggests that the R2D2 protein is the sensor for choosing the guide strand of the siRNA duplex to be loaded into the RISC¹⁵.

To determine whether TRBP may function by bridging the association of Dicer and Ago2 with the siRNA, we assessed the physical association of Ago2 with Dicer–TRBP by mobility shift gel electrophoresis using a radiolabelled siRNA. Although recombinant Dicer did not associate with the siRNA, addition of recombinant Dicer–TRBP (fraction 28 of the Superose 6 gel filtration shown in Fig. 1d) resulted in the formation of a complex between Dicer–TRBP and the siRNA (Fig. 2b). Notably, TRBP alone (fraction 32 of the Superose 6 gel filtration; Fig. 1e) could form a stable, rapidly migrating complex with double-stranded siRNA, which was distinct from that of the Dicer–TRBP complex (Fig. 2c). Although Ago2 cannot associate with the siRNA on its own (Fig. 2d, lane 4), addition of recombinant Ago2 to the Dicer–TRBP–siRNA complex resulted in the appearance of a slower migrating Dicer–TRBP–Ago2 complex (Fig. 2b). These results indicate that TRBP may have a role in

mediating Dicer association with siRNA and the recruitment of Ago2.

Analysis of the interaction between the endogenous Dicer–TRBP complex and the siRNA resulted in the appearance of two complexes: a minor, slower migrating complex, and a major, faster migrating complex (Fig. 2d). Addition of ATP to the reaction did not affect either complex (data not shown). Whereas the faster migrating complex corresponds to that observed with recombinant Dicer–TRBP, the slower migrating complex resembled that of Dicer–TRBP–Ago2. Indeed, addition of increasing concentrations of recombinant Ago2 to native Dicer–TRBP complex resulted in the formation of a slower migrating complex indistinguishable from the minor band observed with the native Dicer–TRBP complex (Fig. 2d). These results suggested the possibility that the Ago2 protein may form a stable complex with Dicer–TRBP *in vivo*.

To test this possibility, we analysed the Dicer–TRBP complex purified through Flag–TRBP, and fractionated by gel filtration for the presence of Ago2. This revealed the precise co-elution of Ago2 with TRBP and Dicer (Fig. 3a). To demonstrate this association

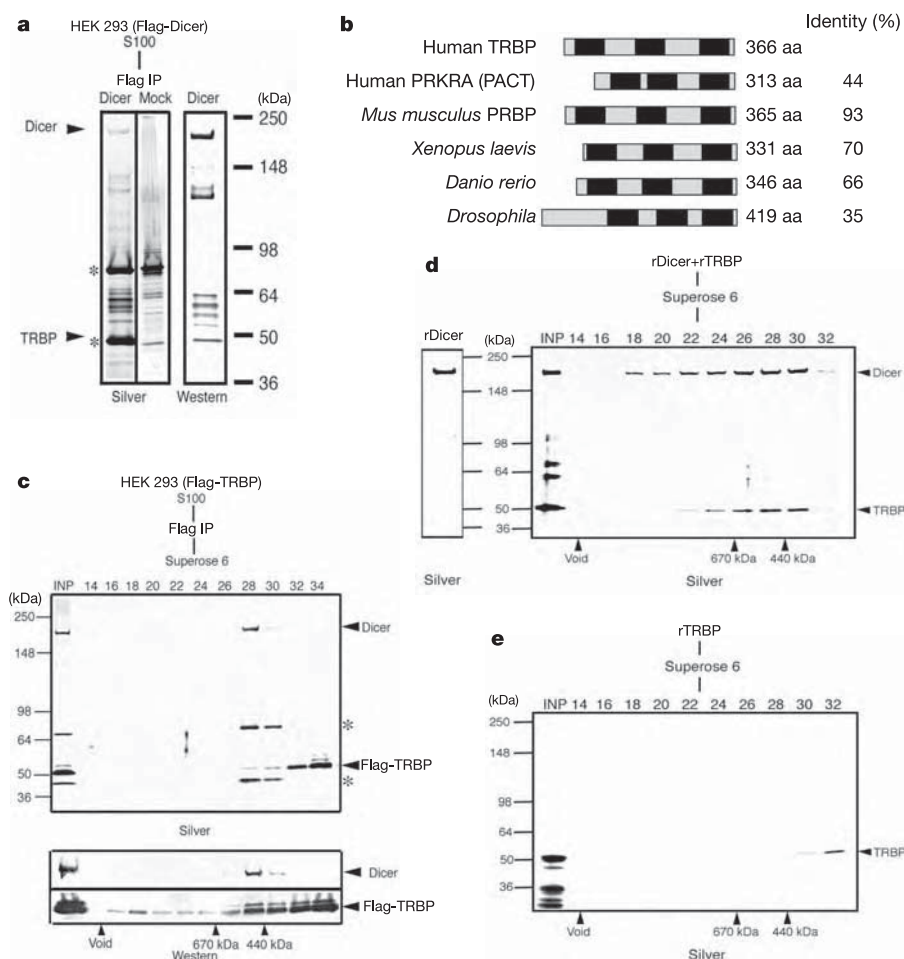


Figure 1 | Isolation of a Dicer–TRBP-containing complex. **a**, Flag–Dicer isolated from a HEK 293-derived cytoplasmic extract (S100) resolved by SDS–PAGE and visualized by silver staining (left panel), and western blot analysis using anti-Flag antibodies (right panel). Arrows point to the bands containing Dicer and TRBP proteins. The molecular masses of marker proteins (kDa) are indicated (right). Asterisks denote contaminating polypeptides that were also present in control immunoprecipitation (IP) from naive cells (Mock). **b**, Diagrammatic representation of the human TRBP protein and alignment with orthologous proteins. Conserved dsRBDs are depicted as black squares. Protein accession numbers (NCBI) are: AAB50581 (human TRBP), NP_003681 (human PRKRA), NP_03345

(mouse PRBP), AAH55390 (*Danio rerio*), MGC82499 (*Xenopus*), NP_609646 (*Drosophila*). aa, amino acids. **c**, The Dicer–TRBP complex isolated from S100 by Flag–TRBP, fractionated by Superose 6 gel filtration, and visualized by silver staining and western blot analysis. INP denotes the input to the column; asterisks denote the common contaminants (SKB1 and MEP50). Fractions of the column are shown along the top; molecular mass markers are indicated along the bottom. **d**, Silver stains of recombinant Dicer (rDicer) and recombinant Dicer–TRBP complex (rDicer + rTRBP) reconstituted and purified by gel filtration. **e**, Silver stain analysis of recombinant TRBP fractionated by gel filtration.

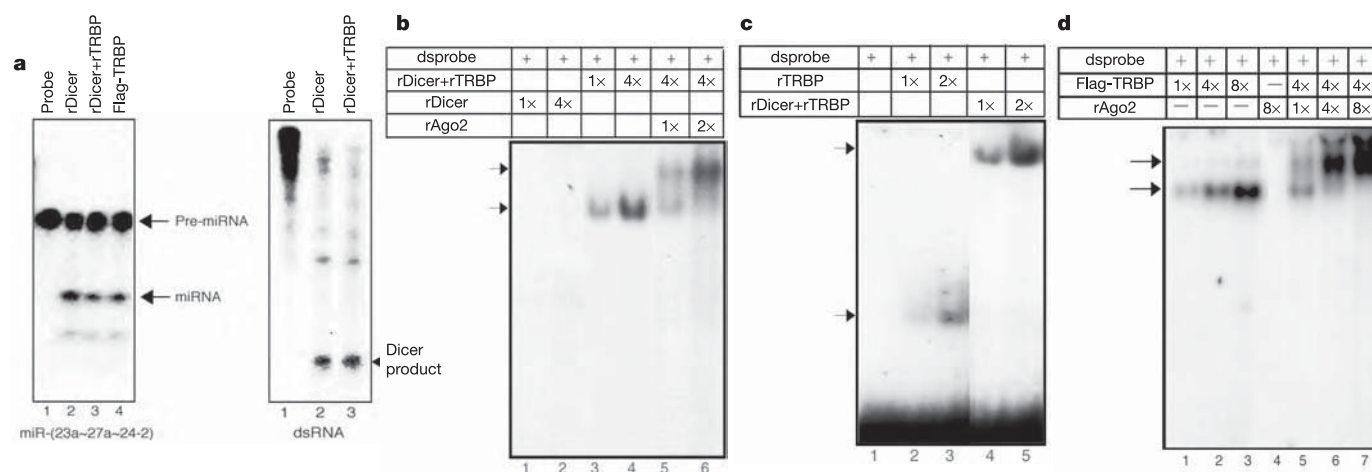


Figure 2 | The Dicer-TRBP complex processes miRNAs and associates with siRNA to form a ternary complex with Ago2. **a**, Pre-miRNA processing assay (left panel). The pre-miRNAs generated by the Microprocessor complex were subjected to a second step of processing using recombinant Dicer, recombinant Dicer-TRBP, or fraction 28 of the Superose 6 column of immunopurified TRBP complex (Flag-TRBP) to yield a 22-nucleotide mature miRNA. The right panel shows the long double-stranded RNA processing assay. Double-stranded RNA was incubated with recombinant Dicer or recombinant Dicer-TRBP to yield a 22-nucleotide mature miRNA. **b**, EMSA performed using radiolabelled double-stranded siRNA. Recombinant Dicer and recombinant Dicer-TRBP (fraction 28 of the

Superose 6 column) was analysed for complex formation with siRNA. 1× corresponds to 20 ng of Dicer and 10 ng of recombinant Ago2. **c**, EMSA performed using radiolabelled siRNA. Recombinant Dicer-TRBP (fraction 28 of the Superose 6 column) and recombinant TRBP (fraction 32 of the Superose 6 column) were analysed for complex formation with siRNA. 1× corresponds to 20 ng for each protein. **d**, EMSA performed with Dicer-TRBP (fraction 28 of the Superose 6 column corresponding to Flag-TRBP) and addition of recombinant Ago2. 1× corresponds to 5 ng of native complex and 2.5 ng of recombinant Ago2. Arrows indicate the two ribonucleoprotein complexes.

rigorously, we generated HEK 293-derived stable lines expressing Flag-Ago2. Flag-Ago2 was purified and subjected to gel filtration. Western blot analysis of Superose 6 gel filtration column fractions revealed the presence of Ago2 in multiple complexes of variable size (Fig. 3b). However, a distinct complex containing TRBP and Dicer was detected peaking in fractions 26–28, consistent with the presence of a Dicer-TRBP-Ago2 complex displaying a similar molecular mass as that observed after analysis of Flag-TRBP by gel filtration (Fig. 3a). Consistent with the presence of multiple isoforms of TRBP, we detected multiple TRBP bands associating with Dicer-Ago2 (Fig. 3b). Taken together, these data reveal the stable association of a trimeric protein complex containing Dicer-TRBP and Ago2 that can assemble onto siRNA duplexes.

To address the role of the Dicer-TRBP complex in the generation of mature miRNAs *in vivo*, we used RNA interference (RNAi) to deplete Dicer and TRBP. Two different siRNAs targeting Dicer and TRBP were used to knockdown the levels of each protein, whereas a siRNA targeting TFII-I was used as a control. Polymerase chain reaction with reverse transcription (RT-PCR) demonstrated the gene specificity of the siRNAs in reducing mRNA levels (Fig. 4a); however, depletion of either Dicer or TRBP resulted in the destabilization of both proteins, reflecting a role for their association in maintaining protein stability (Fig. 4b). Predictably, analysis of miRNA levels after abrogation of Dicer-TRBP resulted in a substantial decrease in the level of mature miRNA (Fig. 4c). The low level of miRNAs detected after Dicer depletion could be due to either residual Dicer protein after siRNA treatment, or may represent persistence of mature miRNAs over the relatively short time period of these experiments.

The association of the Dicer-TRBP complex with Ago2 is suggestive of a coupling of the initiation step of RNAi and its effector role in RNA-mediated gene silencing. We depleted Dicer or TRBP using siRNA and then assessed the extent of post-transcriptional gene silencing by measuring the effectiveness of siRNA targeting firefly luciferase to decrease firefly luciferase activity. A *Renilla* luciferase that was not a target of the siRNA was used as a control to normalize the luciferase activity. Treatment of HeLa cells with siRNA against firefly luciferase resulted in 80% inhibition of the firefly luciferase

activity (Fig. 4d). Depletion of Dicer or TRBP (approximately 80% depletion of Dicer and TRBP levels in three independent experiments) using siRNA relieved this inhibition to about 40% compared with cells that were treated with control siRNA against TFII-I. These

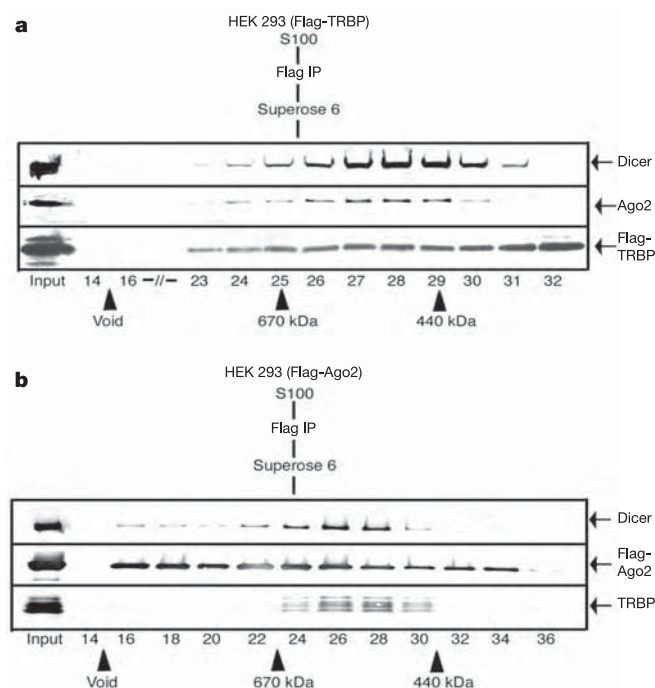


Figure 3 | Stable association of Dicer-TRBP with Ago2. **a**, Western blot analysis of Flag-TRBP affinity eluate fractionated by Superose 6 gel filtration. Column fractions were analysed using the antibodies indicated. **b**, Western blot analysis of Flag-Ago2 affinity eluate fractionated by Superose 6 gel filtration. Column fractions were analysed using the antibodies indicated.

results implicate a role for the Dicer–TRBP complex in linking the initiation step of siRNA production to the execution step of post-transcriptional gene silencing.

We have isolated a trimeric protein complex of approximately 500 kDa that contains TRBP, Dicer and Ago2. We demonstrate that this complex is required for miRNA biogenesis and links miRNA processing with the assembly of an active RISC. This contention is based on the following. First, affinity purification and gel filtration

of Flag–TRBP and Flag–Ago2 revealed a stable association of Dicer–TRBP–Ago2 in a discrete complex of about 500 kDa that displayed pre-miRNA processing activity *in vitro*. Second, the close association of Dicer and TRBP was further demonstrated by the fact that depletion of Dicer or TRBP resulted in the loss of stability of the reciprocal binding partner *in vivo* and concomitant loss of miRNA processing. Third, TRBP was required for the recruitment of Ago2 to the siRNA duplexes. Finally, knockdown of Dicer–TRBP resulted in impaired post-transcriptional gene silencing *in vivo*. Previous studies have pointed to a mechanistic coupling between the initiation phase of RNAi and the execution phase⁹. Specifically, these results point to a role for Dicer in facilitating siRNA-directed mRNA cleavage^{16,17}. Our results extend these contentions and delineate a role for the double-stranded RNA-binding protein TRBP in coupling the initiation and execution steps of RNAi. These results are also consistent with a step-wise assembly of RISC in which the Dicer–TRBP complex, by virtue of its affinity for siRNA, recruits Ago2 to siRNA, leading to the formation of the trimeric complex. This trimeric complex provides the foundation for the assembly of an active holo-RISC complex through the recruitment of additional proteins⁹.

TRBP binds the HIV TAR, a stem-loop structure required for the transcriptional activation of HIV¹⁰. Furthermore, TAR has secondary structure that resembles miRNA precursors, raising the possibility that TAR is a viral pre-miRNA. Moreover, deletion of PRBP, the mouse homologue of TRBP, yields viable mice that often die at weaning. Surviving homozygous mutant males show defects in spermatogenesis¹⁸. In contrast, knocking out Dicer results in embryonic lethality¹⁹. Given that in humans Dicer requires TRBP for protein stability and RISC assembly, it seems likely that in mice there is a redundant factor, possibly PRKRA (also known as PACT in human), that compensates for PRBP during development but not spermatogenesis.

METHODS

Affinity purification of Flag–Dicer, Flag–TRBP and Flag–Ago2. Expression plasmids for Flag–Dicer (carboxy-terminal tag), Flag–TRBP (amino-terminal tag) and Flag–Ago2 (amino-terminal tag) were individually co-transfected with a puromycin resistance plasmid into HEK 293 cells. Complexes were purified (50–150 mg) from cytoplasmic extract (S100) with anti-Flag M2 affinity gel (Sigma). After washing with buffer A twice (20 mM Tris–HCl (pH 7.9), 0.5 M KCl, 10% glycerol, 1 mM EDTA, 5 mM dithiothreitol (DTT), 0.2 mM phenylmethylsulphonyl fluoride (PMSF) 0.5% NP40), and once with buffer B (20 mM Tris–HCl (pH 7.9), 0.1 M KCl, 10% glycerol, 1 mM EDTA, 5 mM DTT, 0.2 mM PMSF), the affinity column was eluted with Flag peptide. Analysis of Flag–TRBP and Flag–Ago2 by gel filtration was as described previously². Protein identification using liquid chromatography tandem mass spectrometry (LC–MS/MS) was performed as described previously². Anti-Dicer, anti-TRBP and anti-Ago2 polyclonal antibodies were generated to the N and C terminus of each protein (Open Biosystems). Recombinant Dicer, TRBP and Ago2 were purified according to methods previously described for purification of recombinant proteins from insect cells and bacteria².

Pre-miRNA and long double-stranded RNA processing. Processing of *in vitro*-transcribed primary miRNA by the Microprocessor complex (Drosha–DGCR8) was carried out as described previously². Long double-stranded RNA was prepared by *in vitro* transcription of a 150-nucleotide RNA from both strands of a DNA plasmid. The two complementary RNAs were mixed together in equimolar amounts in RNase-free water with 10 U RNasin and incubated overnight at 37 °C. After gel purification, long double-stranded RNA was used in processing assays. Processing of pre-miRNA or long double-stranded RNA was performed by incubating the Dicer–TRBP complex, recombinant Dicer, or recombinant Dicer–TRBP complex with gel-purified pre-miRNA in a buffer containing 3.2 mM MgCl₂, 1 mM ATP, 20 mM creatine phosphate, 1 U μl^{−1} HPRI (Takara), 20 mM Tris–HCl (pH 7.9), 0.1 M KCl, 10% glycerol, 5 mM DTT, 0.2 mM PMSF, for 1 h at 37 °C. Subsequent to phenol:chloroform extraction and RNA precipitation, samples were resolved by 15% denaturing polyacrylamide gel.

Electrophoretic mobility shift analysis. Electrophoretic mobility shift analysis (EMSA) was performed using [γ -³²P]ATP end-labelled, double-stranded siRNA (luciferase) as a probe. Protein complexes/recombinant proteins were added to the reaction containing 15 nM siRNA in the same buffer used for pre-miRNA

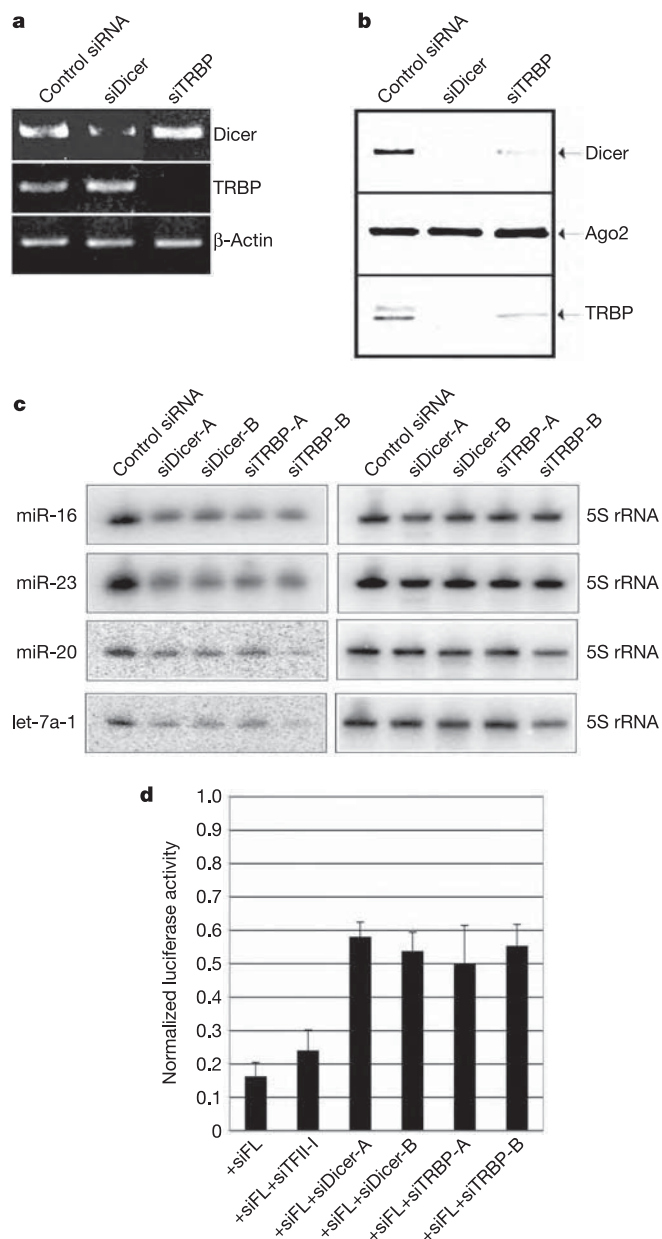


Figure 4 | The Dicer–TRBP complex is required for miRNA biogenesis and post-transcriptional gene silencing. **a**, Analysis of the transcript levels of TRBP, Dicer and β -actin by RT–PCR, after treatment of HeLa cells with a representative siRNA to TFII-I (control), TRBP and Dicer. **b**, Western blot analysis of Flag–Ago2 immunoprecipitate after siRNA-mediated knockdown of TFII-I (control), TRBP and Dicer. **c**, Northern blot analysis of miR-16, miR-23, let-7a-1 and miR-20 after treatment of HeLa cells with the indicated siRNA. **d**, HeLa cells co-transfected with firefly and *Renilla* luciferase plasmids and a combination of siRNA targeting firefly luciferase and one of the siRNAs shown in the bottom of the figure. Firefly luciferase activity was normalized relative to that of *Renilla* luciferase. Error bars represent standard error of the mean for three independent experiments.

processing, and incubated at 30 °C for 20 min. Complexes were resolved by 4% native PAGE at 4 °C.

siRNAs and northern blot analysis. The sequences of the siRNAs used in this study are: Dicer-A, 5'-UUUGUUGCGAGGCGUGAUUCdTdT-3'; Dicer-B, 5'-UCUAAUAGCACCUGAUGUdTdT-3'; TRBP-A, 5'-GCUGCCUAGUAUAGAGCAAdTdT-3'; TRBP-B, 5'-UCUACGAAAUUCAGUAGGAdTdT-3'; Luciferase, 5'-UCGAAGUAUCCGCGUACGdTdT-3'; TFII-I, 5'-UGUGGGAAGCUCUUGGCCdTdT-3'. Transfection of HeLa 293 cells was performed with Lipofectamine 2000 (Invitrogen), as described². To examine the effect of siRNA on target gene expression, total RNA and complementary DNA synthesis was prepared 48–72 h after transfection, as described². Primer sequences for RT-PCRs in Fig. 4a are: Dicer, 5'-CATGGATAGTGGGATGTCAC-3' and 5'-CTACTTCCA-CAGTGACTCTG-3'; TRBP, 5'-GGGCTGCCTAGTATAGAGC-3' and 5'-CCCTGACCACTGCAGCTGGTG-3'; β -actin, 5'-AAAGACCTGTACGCCAACAC-3' and 5'-GTCATACTCCTGCTTGCTGAT-3'. For northern blot analysis, total RNA (5–10 μ g of each sample, isolated using TRIzol reagent) was resolved on 15% denaturing polyacrylamide gel and electrotransferred onto Hybond N+ nylon membranes (Amersham). Hybridization with specific [γ -³²P]ATP end-labelled DNA oligonucleotides was carried out as described previously².

Luciferase assay. Luciferase assays were performed as described¹⁷ with the following modification: HeLa cells were first co-transfected with firefly and *Renilla* luciferase reporter gene expression plasmids. A second transfection of siRNA against TRBP, Dicer or TFII-I together with siRNA against firefly luciferase was performed.

Received 22 March; accepted 3 June 2005.

Published online 22 June 2005.

1. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
2. Gregory, R. I. *et al.* The Microprocessor complex mediates the genesis of microRNAs. *Nature* **432**, 235–240 (2004).
3. Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F. & Hannon, G. J. Processing of primary microRNAs by the Microprocessor complex. *Nature* **432**, 231–235 (2004).
4. Han, J. *et al.* The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev.* **18**, 3016–3027 (2004).
5. Bernstein, E., Caudy, A. A., Hammond, S. M. & Hannon, G. J. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**, 363–366 (2001).
6. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**, 23–34 (2001).
7. Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834–838 (2001).
8. Ketting, R. F. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes Dev.* **15**, 2654–2659 (2001).
9. Sontheimer, E. J. Assembly and function of RNA silencing complexes. *Nature Rev. Mol. Cell Biol.* **6**, 127–138 (2005).
10. Gagnol, A., Buckler-White, A., Berkhout, B. & Jeang, K. T. Characterization of a human TAR RNA-binding protein that activates the HIV-1 LTR. *Science* **251**, 1597–1600 (1991).
11. Liu, J. *et al.* Argonaute2 is the catalytic engine of mammalian RNAi. *Science* **305**, 1437–1441 (2004).
12. Meister, G. *et al.* Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol. Cell* **15**, 185–197 (2004).
13. Zhang, H., Kolb, F. A., Jaskiewicz, L., Westhof, E. & Filipowicz, W. Single processing center models for human Dicer and bacterial RNase III. *Cell* **118**, 57–68 (2004).
14. Liu, Q. *et al.* R2D2, a bridge between the initiation and effector steps of the *Drosophila* RNAi pathway. *Science* **301**, 1921–1925 (2003).
15. Tomari, Y., Matranga, C., Haley, B., Martinez, N. & Zamore, P. D. A protein sensor for siRNA asymmetry. *Science* **306**, 1377–1380 (2004).
16. Lee, Y. S. *et al.* Distinct roles for *Drosophila* Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell* **117**, 69–81 (2004).
17. Doi, N. *et al.* Short-interfering-RNA-mediated gene silencing in mammalian cells requires Dicer and eIF2C translation initiation factors. *Curr. Biol.* **13**, 41–46 (2003).
18. Zhong, J., Peters, A. H., Lee, K. & Braun, R. E. A double-stranded RNA binding protein required for activation of repressed messages in mammalian germ cells. *Nature Genet.* **22**, 171–174 (1999).
19. Bernstein, E. *et al.* Dicer is essential for mouse development. *Nature Genet.* **35**, 215–217 (2003).

Acknowledgements We thank T. Beer for assistance with the microcapillary HPLC/mass spectrometry. R.S. was supported by a grant from NIH and American Cancer Society. R.I.G. is a fellow of the Jane Coffin Childs Fund for Medical Research.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to R.S. (shiekhattar@wistar.org).

-  FOCUS
-  SPOTLIGHT
-  RECRUITMENT
-  ANNOUNCEMENTS
-  EVENTS

naturejobs

You've got to laugh...

When Jorge Cham started his PhD at Stanford in 1997, he felt the pressure familiar to most graduate students. But he dealt with it a bit differently — by starting his own comic strip. Cham says that the strip, called *Piled Higher and Deeper*, helped him to process his own “stress and anxiety”.

It started out as doodles inspired by *Archie*, the perpetual comic teenager from the 1950s, and *Doonesbury*, the political satire that infamously featured US President George W. Bush as a talking feather sporting a cowboy hat. But it soon turned into a vehicle for empathy. “No one was really telling the story of the grad student,” Cham says. Word spread as Cham uploaded the cartoon to the Internet during the embryonic stage of the web, when there was a great need for content, traffic was growing and there was little competition.

The unexpected impact of the strip connected Cham to other graduate students who were experiencing the same crisis of confidence, crowded lab conditions and long hours that had initially overwhelmed him. The strip was helping to ease the stress of readers and creator alike.

Cham says that most of the comments he gets fall into four categories: a thank-you for adding to their procrastination rituals; mock confusion over whether they should laugh or cry; a sometimes disturbing identification with one of the recurring characters; and the revelation that they no longer feel alone in being mired in grad-school stress.

Even though Cham has evolved from Stanford graduate student to California Institute of Technology neuroscientist instructor, he is still penning and posting the strip. He celebrated the 600th this year and published his second collection of cartoons in May. *Piled Higher and Deeper* will now be available on the Naturejobs graduate channel — where readers are free to use them for procrastination or decompression, as they see fit.



Paul Smaglik, Naturejobs editor

CONTACTS

Publisher: Ben Crowe
Editor: Paul Smaglik
Marketing Manager: David Bowen

European Head Office, London
 The Macmillan Building, 4 Crinan Street
 London N1 9XW, UK
 Tel: +44 (0) 20 7843 4961
 Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com

Naturejobs Sales Director:
 Nevin Bayoumi (4978)
European Sales Manager:
 Andy Douglas (4975)

Natureevents: Sille Opstrup (4994)
UK/RoW/Ireland/Italy:
 Nils Moeller (4953)
 Irene Viglia-Atton (4944)
Scandinavia/Spain/Portugal:
 Evelina Rubio Håkansson (4973)
France/Switzerland/Belgium:
 Amelie Pequignot (4974)
Germany/Austria/The Netherlands:
 Reya Silao (4970)

Advertising Production Manager:
 Billie Franklin
 To send materials use London
 address above.
 Tel: +44 (0) 20 7843 4814

Fax: +44 (0) 20 7843 4996
 e-mail: naturejobs@nature.com
Naturejobs web development: Tom Hancock
Naturejobs online production: Niamh Shields

European Satellite Office
Germany/Austria/
The Netherlands:
 Patrick Phelan
 Tel: +49 89 54 90 57 11
 Fax: +49 89 54 90 57 20
 e-mail: p.phelan@nature.com

US Head Office, New York
 345 Park Avenue South,
 10th Floor, New York,

NY 10010-1707
 Tel: +1 800 989 7718
 Fax: +1 800 989 7103
 e-mail: naturejobs@natureny.com

US Sales Manager: Peter Bless

Japan Head Office, Tokyo
 Chiyoda Building,
 2-37 Ichigayatamachi,
 Shinjuku-ku,
 Tokyo 162-0843
 Tel: +81 3 3267 8751
 Fax: +81 3 3267 8746
Asia-Pacific Sales Director: Rinoko Asami
 e-mail: rasami@naturejpn.com

An individual approach

Reduced side effects and more effective therapies are some of the benefits promised by pharmacogenomics. But to reach these goals industry will have to marshal a broad range of skills, as **Ricki Lewis** explains.

It would be the ultimate in personal service: medical treatment tailored to match exactly your personal genetic make-up. Understanding how your body is likely to react to certain therapies would potentially eliminate side effects and increase the efficacy of the drugs you take. That, loosely, is the promise of pharmacogenomics — but until recently it seemed a very distant goal.

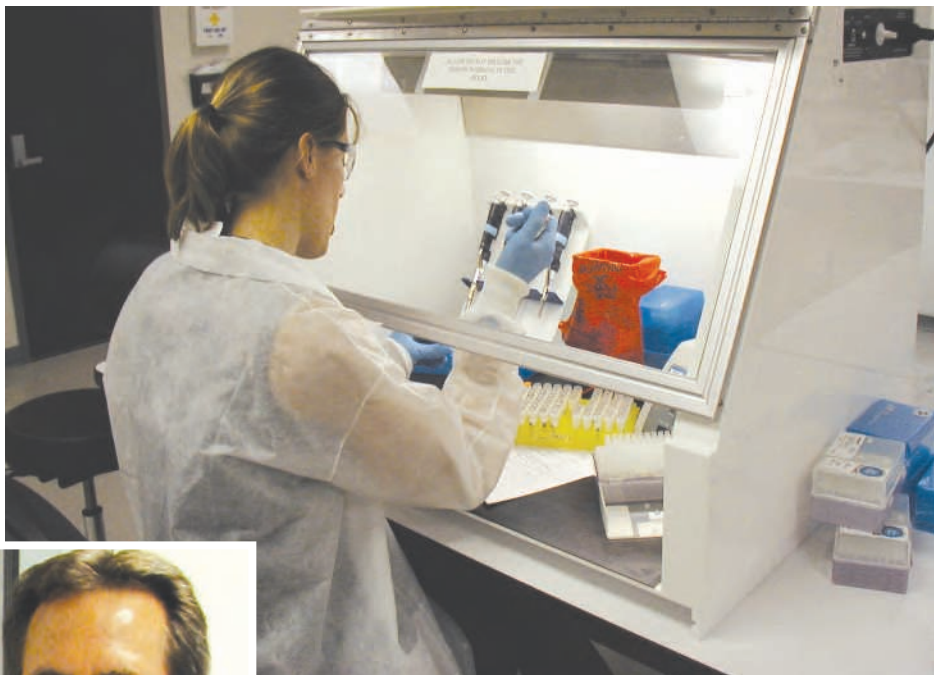
Events last autumn have changed all that. When a group of painkillers known as COX-2 inhibitors, used to treat arthritis, were linked to an increased risk of heart disease, the idea of personalized medicine gained some very public momentum (see 'A long time coming', opposite). The idea that targeting drugs to individuals can save lives has helped to boost the role of pharmacogenomics.

In March, the US Food and Drug Administration (FDA) released guidelines that encourage, but do not require, companies to provide pharmacogenomic data in new applications for drug approval. This information would predict how a patient might respond to a drug based on their genetic profile: both the gene variants they have inherited and the expression pattern of those genes. This small regulatory step will translate into jobs all along the drug discovery and development pipeline. Those who can acquire the right mix of skills now will be poised to take advantage of what some see as a complete restructuring of the pharmaceutical industry.

"The FDA guidance document has opened up a large opportunity for people with pharmacogenomics knowledge to help pharmaceutical companies navigate the path forwards," says Michael Murphy, president and chief executive of Gentriss, a pharmacogenomics company in Morrisville, North Carolina. Gentriss has already opened a consulting division to meet the growing demand for expertise in combining genomics knowledge with statistics, bioinformatics and quality assurance.

In the past five years, every large pharmaceutical company has set up a pharmacogenomics group, says John Ryan, senior vice-president of translational medicine at Wyeth Pharmaceuticals in Collegeville, Pennsylvania. "Based on the technology, the resurgence of biotech, and the fact that the drug industry is being pressed to do things differently, there will be hundreds of openings in pharmacogenomics over the next several years," he predicts.

Wolfgang Sadée, director of the pharmacogenomics programme at Ohio State University in Columbus, adds that it will be five years until drug companies fully



Michael Murphy believes that pharmacogenomics will offer many career opportunities.

Peering under the hood: a deeper understanding of gene expression will help tailor therapies to individual needs.

GENTRIS

embrace the field. "If we train people now," he says, "in five years they will be in the right position to help out."

The earlier approach of pharmacogenetics, which targets a single gene, has already scored some successes. Herceptin (trastuzumab), for example, treats breast cancer by targeting a single receptor on tumour cells that is overabundant in women who inherit a specific gene mutation. And the analysis of individual genes that affect drug metabolism has helped to improve efficacy.

With its emphasis on multiple genes and their expression patterns, pharmacogenomics extends the reach of its predecessor. For instance, researchers at the University of Toronto in Canada have identified 18 genes whose expression pattern predicts whether or not a person will respond to treatment for hepatitis C. Current treatment (interferon plus ribavirin) works for only half of them. Measuring expression of these genes could indicate who can be helped, right from the start.

A mix of skills

Pharmacogenomics requires expertise in genetics and genomics, drugs and drug interactions, bioinformatics and especially statistics, says Kelly Frazer, vice-president of genomics at Perlegen Sciences, a pharmacogenomics company in Mountain View, California. The ability to communicate is also essential, she stresses: people work in teams, at least in the short term, because not many individuals have all the necessary skills.

The field leans heavily on two techniques: analysis of SNPs (single nucleotide polymorphisms) and DNA microarrays. SNPs are sites in the genome where the DNA base differs in at least 1% of a population. Sets of SNPs along chromosomes form haplotypes, which are the chunks of information associated with specific

drug responses. DNA microarrays are used to identify inherited SNP patterns — a singly array can screen 100,000 SNPs in just hours. Microarrays that represent the mRNAs in a particular cell type are used to track gene-expression patterns that can correlate to drug response.

The real work to come will be in data-mining. “There is still a gap — but it is shrinking — in the ability to understand the data from a genomics perspective, manipulate it from a computational perspective, and analyse it from a statistical perspective,” Frazer adds. Audrey Papp, manager of the pharmacogenomics lab at Ohio State University, agrees: “We need people to look at the information and pull it all together. How do the data fit? Which data are significant?”

Learning the ropes

Several trajectories lead to the skill sets of a pharmacogenomics specialist. David Gurwitz, director of the National Laboratory for the Genetics of Israeli Populations at Tel Aviv University, suggests taking pharmacogenomics courses in medical, pharmacy or health-science schools, and seeking a mentor who is actively working on a pharmacogenomics problem.

Doctoral degrees in pharmacy or clinical pharmacology provide excellent preparation, says Murphy. Demand for Ohio State University's pharmacogenomics programme is high, adds Sadee, who directs the programme, which was set up three years ago in the department of pharmacology. The programme draws students from an integrated biomedical graduate programme, neuroscience and biomedical informatics. “Quite a number of young people are now interested in the area. I get e-mails all the time from people in academic departments requiring this type of expertise,” Sadee says.

A degree in biostatistics is another route to pharmacogenomics. “The area where most of the jobs will be is in bioinformatics, and this is where the industry is the least well-served. Biostatisticians specifically trained in pharmacogenomics is a growth area,” says Ryan.

Education can also stem from conferences — such as the Cold Spring Harbor–Wellcome Trust meeting this September in Hinxton, Cambridge, UK — or through industry. Celera Genomics of Rockville, Maryland, for example, offers intern opportunities and recruits recent college graduates.

Job opportunities range from the basic to the applied, and from tightly focused tasks to broader responsibilities. Industry, for example, requires technicians to sequence DNA, and study directors to oversee the entire journey of correlating genotypes and expression patterns to drug responses.

“A study director has basic scientific knowledge and knows how to do quality-assurance checks on genomic data,” explains Murphy. Gentris has had to recruit externally to fill these positions, and has set up in-house training for internal promotion, where professionals with RQAP-GLP (registered quality assurance professional in good laboratory practices) certification teach traditionally trained scientists the regulatory ropes.

Another niche is in hospitals, where adverse drug reactions attest to the need for the field. It is estimated that there are about two million adverse drug events in the United States a year, of which about 100,000 are fatal. “As pharmacogenomics slowly moves into the

A LONG TIME COMING

The recent problems surrounding the COX-2 group of painkillers illustrate how pharmacogenomics could in future limit side effects. After revelations last year that extended use of some COX-2 inhibitors could increase the risk of heart problems, action was swift. Merck voluntarily took its COX-2 inhibitor Vioxx (rofecoxib) off the market in September 2004. The licence for Pfizer's Bextra (valdecoxib) was later withdrawn by the US Food and Drug Administration, although the company's Celebrex (celecoxib) was allowed to stay on sale subject to certain restrictions.

While Merck and Pfizer were worrying about the loss of their blockbuster drugs, a test for gene expression was offering a glimpse of the future. The test, introduced by Roche Diagnostics in 2004, detects variants of the genes that encode the proteins cytochrome P450, 2D6 and 2C19, which regulate the rate at which the liver metabolizes many drugs. Patients identified as ‘poor metabolizers’ risk overdosing on what for others would be standard doses.

Not only will pharmacogenomics become standard in future drug discovery: it is already fine-tuning use of old drugs. Consider the anticoagulant warfarin, used by millions since the 1940s to prevent blood clots after major surgery, stroke or heart attack. Matching dose to patient is notoriously tricky, and if it's misjudged a patient could either bleed to death or suffer a fatal blood clot. A team from the National Institute of General Medical Sciences' Pharmacogenetics Research Network recently identified gene variants that correlate to requiring low, moderate or high doses of this commonly prescribed drug.

R.L.



Wolfgang Sadee sees training now as the key to future success within personalized medicine.

health-care arena, more and more hospitals are going to set up teams to monitor adverse events due to drugs. Hospitals will need people to consult on what happens to those patients,” says Murphy.

Teaching opportunities will also span the lab and the clinic. “There will be an urgent need for many pharmacogenomics instructors, both for professional education and for explaining to patients the meaning of tests and their results,” Gurwitz predicts.

Government research groups and consortia offer opportunities, too, such as the Pharmacogenetics Research Network of the the National Institute of General Medical Sciences (NIGMS), and its knowledge base to handle the incoming data, PharmGKB. The International HapMap Project is linking millions of SNPs to drug responses at ten centres around the world, and introduces yet another job title: ‘community engagement and sample collection’. This involves acquiring DNA samples from populations in the United States, China, Japan and Nigeria.

The extensive partnering between big drug companies and smaller biotech firms in pharmacogenomics presents opportunities for researchers in one sector to explore the other. For example, Eli Lilly of Indianapolis and ParAllele BioScience of South San Francisco are analysing 180 genes and their 2,000 variants that affect drug metabolism, using ParAllele's microarray system. And Roche, based in Basel, Switzerland, investigates the pharmacogenomics of asthma and high blood pressure with deCODE Genetics of Iceland.

The future job market in pharmacogenomics is intimately tied to the gestation of the field. The first phase of personalized medicine, says Ryan, is already under way, with the genotyping required to prescribe Herceptin for breast-cancer patients. But there's still a long way to go.

“The end result — each individual having a genomic profile at the doctor's office for every drug — is decades away,” he adds. In the meantime, those who can read drug responses in DNA sequences will enjoy an increasingly robust job market.

Ricki Lewis is a freelance writer in Scotia, New York.

WEB LINKS

FDA pharmacogenomics guidelines

♦ www.fda.gov/cder/genomics/default.htm

Pharmacogenomics knowledge base

♦ www.pharmgkb.org

International HapMap project

♦ www.hapmap.org

SNP database

www.ncbi.nlm.nih.gov/projects/SNP

Pharmacogenetics research network

♦ www.nigms.nih.gov/pharmacogenetics

On firm foundations

Flexible and relatively unfettered, non-profit foundations are able to boldly go into areas of research funding often untouched by public bodies, says **Helen Gavaghan**.

Non-profit foundations that fund research are frequently seen as harbouring huge endowments and dishing out nearly unlimited resources. But the numbers tell a different tale.

Although the ten biggest grant-giving scientific foundations have endowments totalling over \$65 billion, their yearly grant total is a mere fraction of that — about \$3 billion. The figures pale even further when set against the amount governments spend on science and technology. A single US agency, the National Institutes of Health, boasts an annual budget of more than \$28 billion, directed primarily at research.

Foundations do, however, hugely influence the kind of science that gets done, where it takes place and how it is conducted. Because they target research that governments might be avoiding at the time — especially projects seen as too interdisciplinary to gain the support needed, or as premature, politically risky or simply unpopular — foundations can also have a profound effect on individual careers.

In the United States, the New York City-based Alfred P. Sloan Foundation helped launch the Census of Marine Life, which aims to record the diversity, distribution and abundance of life in the oceans. After the Sloan's initial funding, the project drew in a number of other foundations and government agencies. In Britain, the Wellcome Trust — which played a major role in funding the Human Genome Project — helped establish the 'biobank' project, a database for exploring how genetics, environment and diet influence disease. Both projects have resulted in funding for hundreds of

jobs, and the foundations' leaders have also drawn in government money to create even more posts.

That influence is why the science community closely follows the moves of the Wellcome Trust and its fellow members in the Big Three of biomedical research — the Bill & Melinda Gates Foundation in Seattle, Washington, and the Maryland-based Howard Hughes Medical Institute (HHMI). For specialists, smaller foundations focusing on a few key issues or diseases are also worth watching, as they can put a disproportionate amount of resources into a highly specific area.

All change

The Big Three have all been seriously busy of late. Perhaps most significantly, the Gates Foundation, founded by Microsoft chairman Bill Gates, is transforming itself from an organization that pays for vaccinations of known diseases in the developing world to one that also funds primary research through programmes such as its Grand Challenges in Global Health initiative. The HHMI, meanwhile, is building and staffing its own \$500 million research facility in northern Virginia, dubbed Janelia Farm.

Foundations tend to move fast when the need for change arises because they are free from the limitations of government policy. "As a charitable or independent funder, we don't have to go with the political flow," says Robert Terry, an analyst at the Wellcome Trust. In 2004,

for instance, the trust spent \$641 million on biomedical research in Britain, in developing countries such as Zimbabwe, and in Central and Eastern Europe.

And all of the Big Three are granting resources to historically underfunded research on tropical diseases such as malaria, which primarily affect the developing world. For example, the Grand Challenges — part-funded with \$27.1 million from the Wellcome Trust and \$4.5 million from the Canadian Institutes for Health Research — will pay 43 research groups to tackle pressing health issues such as preventing the transmission of malaria and developing new antimalarial drugs. And in July, the HHMI announced a grant of \$17.5 million to fund studies by 42 scientists focusing on the molecular mechanisms of infectious and parasitic diseases.

This flexible response to new areas is a hallmark of foundation funding. Antonio Coutinho, immunologist and director of the Gulbenkian Science Institute in Lisbon, Portugal, has firsthand experience of foundations in France, Switzerland, Sweden and Portugal. What these funders hold in common, he says, is that they "give more freedom to researchers and



All for one: Janelia Farm's Gerald Rubin champions interdisciplinary research.

Top biomedical research funders

Foundation	Assets	Grant making	Subjects funded
Bill & Melinda Gates Foundation	\$26.8 billion	\$1.2 billion*	Development and vaccines
Wellcome Trust	\$18.7 billion	\$642 million†	Biomedicine and history of medicine
Howard Hughes Medical Institute	\$12.8 billion	\$500 million†	Biomedicine
Pasteur Institute	\$570 million‡	\$270 million*	Biomedicine
Gordon and Betty Moore Foundation	\$4.8 billion	\$600 million	Ecosystems, marine biology
Robert Wood Johnson Foundation	\$7.9 billion	\$0.39 million	Health-related, no biomedicine
Juvenile Diabetes Research Foundation	\$100 million‡	\$120 million	Basic and applied research on diabetes
Doris Duke Foundation	\$1.5 billion	\$52.6 million	Environment and medical research
Alfred P. Sloan Foundation	\$1.4 billion	\$64.8 million	Census of Marine Life
W. M. Keck Foundation	\$1.3 billion	\$45.4 million	Science, engineering and medicine

*Year ending December 2003; †Financial year 2004; ‡Bulk of funding from annual fundraising and other sources
Source: Foundation Center, New York, and the European Foundation Centre



Trailblazer: Paul Burn of the diabetes research body JDRF breaks funding bottlenecks.

Delving deep: the Howard Hughes Medical Institute's Janelia Farm is designed for flexible, yet focused research.

have a clear vocation to take risks". And one never knows when a new, wealthy donor will endow a specialist foundation, such as the European Champalimaud Foundation (see 'Vision quest', below), launched in June to research vision.

With its emphasis on interdisciplinary research, Janelia Farm takes flexibility even further. Centre director Gerald Rubin, who led research

to sequence the genome of the fruitfly *Drosophila melanogaster*, has designed Janelia's groups to use a disparate array of skills, tools and knowledge to answer focused questions. The first of these will draw on cell biology, neuroscience, computational biology, developmental biology and genetics to look at how the brain's neuronal circuits process information.

The centre is already well on its way. It has appointed seven of its eventual 24 principal investigators. The campus will open in 2006 and be fully staffed by 2009, with 300 posts and space for 100 visiting researchers.

Web links

Wellcome Trust
 ♦ www.wellcome.ac.uk
 Howard Hughes Medical Institute
 ♦ www.hhmi.org
 The Foundation Center
 ♦ <http://fdncenter.org>
 The European Foundation Centre
 ♦ www.efc.be

VISION QUEST

A new prize is offering a Nobel-sized payout solely for vision research. Europe's most recent foundation for science, the Champalimaud, is to offer a €1 million annual award for the most outstanding research anywhere into vision in any scientific discipline.

"The prize will raise the profile of vision research," says Antonio Coutinho, Gulbenkian Science Institute head and a trustee of the foundation. "The research could range from finding a way to eradicate onchocerciasis or river blindness in Africa, to developing methods of mapping to the visual cortex the light that falls on the retina." The World Health Organization estimates that onchocerciasis has left 270,000 people blind and impaired the vision of a further 500,000.

H.G.

There will be no grant writing, teaching or administrative duties to divert attention from research, and on-site childcare will be provided.

But there are very few HHMIs in the world, leaving scores of more specialized foundations to fill in many gaps. For example, the Juvenile Diabetes Research Foundation International, which spent \$120 million on research in 2004, began funding human embryonic stem-cell research before US President George W. Bush authorized the use of federal dollars to fund studies on a few existing stem cell lines. "Whenever we recognize there's a bottleneck, we try to overcome that," says research director Paul Burn. So now that other foundations and some state or regional governments are funding stem-cell research, JDRF is looking to break other bottlenecks — such as the effects of hypoglycaemia on the central nervous system.

Larger foundations are also nimble at changing their research focus, so

researchers need to check carefully whether ongoing projects are still actually disbursing new funds. Take HHMI's three current international programmes. Only one, for biomedical research in Canada and South America, is still open for new grant proposals. And some long-running programmes at the Sloan on theoretical neurobiology and computational molecular biology are no longer awarding new money.

Stamina needed

Getting to grips with the ins and outs of scientific foundation funding is tough for another reason. Differences in national financial reporting rules make information on foundations difficult to extract, collate and compare in a meaningful way, despite efforts to simplify the process by organizations on both sides of the Atlantic — the US-based Foundation Center and the European Foundation Centre. For example, the FC's top ten list of US science foundations failed to include the HHMI because its grant-giving status differs from that of the others listed. Yet in terms of both endowment and grant-giving, it ranks third globally (see Table, page 748).

This sort of disparity means grantees should be cautious. The EFC advises those in pursuit of research euros to search websites for foundations' aims before submitting a grant application — but this will require stamina. The EFC's website contains data from a 2003 survey of some 27,000 foundations in nine European countries. Using 'science' as a search term yielded only 125 organizations, many of which seemed to have little to do with science funding, while notable science funders such as the Gulbenkian Science Institute in Portugal were missing from the list.

The bottom line is that while foundations' budgets may be dwarfed by worldwide government research and development spending, their impact is substantial. For those who have bold proposals that stand little chance of winning public money, finding the right foundation to match their needs may be well worth the effort — as long as they monitor the winds of change.

Helen Gavaghan is a freelance writer based in West Yorkshire, UK.

Pigs on the wing

Aurorae in the sky with diamonds, just \$10.99 (exc. tax).

K. Erik Ziemelis

"Come on, Kirsty. Get a wriggle on!"

Frank has another look at his watch and glances out the window. No panic yet, but they're cutting it fine.

"I'm coming, I'm coming!"

Kirsty (age 6) not so much descends the stairs as free-falls into the living room, a flurry of brightly decorated limbs that seem to make contact with everything but the ground. Forward momentum absorbed by collision with the waiting adult, she gets to her feet, gives her uncle's legs a brief hug and presents herself for inspection.

"Nice one, kiddo. Let's hit the road."

BreathEazys firmly in place, the two scamper through the heat and haze of the forecourt to the carpool. Frank checks the gauge, disconnects the charger and bundles his niece into the waiting vehicle. Sliding in alongside, he seals the doors, gives the air a quick flush and keys in his ID.

"Colour?"

"Green!"

The Urban EcoWarrior shimmers from neutral grey to fluorescent emerald before rolling smoothly out of the lot to join the silent rainbow of evening rush-hour traffic. The two intrepid explorers settle back in their seats.

"So where are we going?"

Before Frank can articulate his well-rehearsed answer (it is supposed to be a surprise after all), a shrill voice pipes up from the vicinity of Kirsty's feet.

"We're off to see a rory who lives in the sky!"

Momentarily taken back by both the unexpected intrusion of a third party and the worrying possibility that his plans have somehow been exposed, Frank looks down at the floor. And there, easing its way out of the confines of Kirsty's carry-all, is a small green pig with a smug expression on its face. Frank grins.

"Mr Loop, I might have guessed. Are any secrets safe from you? But I think for now you've said enough."

The pig takes in Kirsty's questioning look, glances over at a now dramatically glowering Frank, before returning its attention to the child. Offering what passes for a porcine shrug and a forlorn my-lips-are-sealed wrinkling of its snout, the little animatronix bounces on to Kirsty's lap and snuggles down for the ride.

After a painless drive through the 'burbs, they are out in the foothills. Shortly after-

wards, they pull up alongside the transfer station — no magtrak beyond this point.

Timecheck. The sun still hovers over the island of glass and concrete that they have left behind, but the sky has already lost its familiar ochre tint. Sunset is little more than an hour away.

"OK kiddo, wrap up warm."

Frank climbs swiftly into his weather-proof and hands a miniature but more colourful version to an oblivious Kirsty. For the past 20 minutes her face has been pressed up against the window, gazing wide-eyed at the surrounding countryside. And now, free of both the car and



the claustrophobic air filter, she stands transfixed by the row-upon-row of four-wheeled antiques parked all around them. (Mr Loop on the other hand gives the impression of being fast asleep, but — as befits a miniature replica of the comically self-important children's entertainer — is probably still sulking.)

A gust of chill mountain air snaps Kirsty out of her reverie. Thermals donned, backpacks stowed and inert pig pocketed, the two slide into a brilliant-red, open-topped barchetta. Now it is Frank's turn to shiver with anticipation. The last time he drove one of these he was barely out of his teens, yet he still remembers the thrill of opening up the throttle and hearing the throaty roar of a real engine. And roar it does, right on cue, eliciting a squeal of excitement from the diminutive passenger — and the sweet taste of long-forbidden fruit for the driver. From here, the only way is up.

Final destination: Look-See Point. After a jarring drive up rough mountain tracks, Frank parks beside one of the crumbling domed structures scattered across the high plateau. He gathers a couple of folding chairs and strides hurriedly towards some clear ground. Kirsty only just keeps up and

can barely hide her disappointment when she finally settles herself into a chair and looks around.

"Aw, they're just ruins."

"Never mind those old stones, kiddo. I promised to show you the heavens, and I will. But for now just watch...and wait."

A glance at his watch again: ten minutes to go. Frank smiles in relief. The number of favours he had to call in to pull this off — not to mention the cost of getting a prime-time slot — will not have been wasted.

With the sky now darkened to an inky black, Kirsty spots the first tiny pinpoints of light. 'Stars', her uncle had called them. Sure, they seem to map out some pretty shapes, but she still can't see what the fuss is all about. To cover her mounting gloom, she returns to playing with the pig.

Then the heavens explode with splashes of light and colour. One after another, familiar and unfamiliar objects parade across the sky: items of food, items of clothing, items of...well, just about anything a young mind can imagine, and a lot, lot more. Faster and faster the images race by until all that remains is a pulsating emerald sheet. The rippling edges of this

curtain of light then start to bend and twist, turning in on one another, transforming, transfixing, becoming...

"Mr Loop!"

Frank beams down at his niece. Kirsty, now on her feet, has one hand clamped over her mouth, the other outstretched as if trying to grab the sky itself. About her, the animated toy frolics, uttering decidedly un-pig-like yelps as it flings itself repeatedly into the air. And its luminous namesake responds in kind, leaping nimbly from one celestial point to another, and leaving in its wake a glowing contrail that resolves into the words 'HAPPY BIRTHDAY KIRSTY'. Fade to green. Commercial services will resume shortly.

Frank sweeps his now-exhausted niece into his arms and carries her slowly back to the jeep, the miniature Mr Loop trotting contentedly in their wake. But Kirsty's tired eyes remain locked on the heavens as assorted consumer items recommence their endless auroral display.

"Show's over kiddo. That's enough rory for one night."

K. Erik Ziemelis wishes it to be known that he is almost (but not quite) entirely unrelated to the physical sciences editor of *Nature*.

JACEY